NOTES DE COURS DE MDD203, PROBABILITÉS ET STATISTIQUE, 2020

G. David

4 novembre 2021

Table des matières

J'essaie de modifier ces notes provisoires de temps en temps, et de les mettre sur ecampus assez régulièrement. Je ne crois pas que je mettrai des numéros de version. Attention : même si ces notes sont distribuées, il restera des fautes de frappe. Ne jamais croire à fond quelque chose, sous pretexte que c'est imprimé avec des jolies fontes (ou même pas ?). Utilisation prévue au départ : compenser un peu les ennuis dûs à l'enseignement comodal en laissant une trace de ce que je raconterai en cours.

Désolé, c'est relativement peu dense ; du coup j'espère que ca sera facile à lire, et aussi les fautes de frappe devraient être plus faciles à corriger.

Normalement, des exercices en WIMS doivent être en ligne (ou y seront), avec un mode d'emploi disponible et aussi des ressources de rappels de cours. J'espère aussi que le mode d'emploi contient la manière dont la note est calculée. En tout cas, en cas de question, consultez d'abord le forum (vous avez souvent les mêmes questions). Et n'oubliez pas qu'il y a des dates limites pour faire les devoirs qui comptent !

Je rappelle encore que pour WIMS vous devez avoir moyen de vous entraîner d'abord, mais juste un certain nombre d'essais avant que cela compte. Ce genre de chose est modifiable, donc si vous avez des remarques, plaintes, et sugestions à faire, à la fois sur le système et sur les exos, n'hésitez pas à nous en faire part.

Merci à O. Hénard, qui m'a laissé copier sans vergogne ses anciennes notes manuscrites.

Voici un certain nombre de livres qui peuvent vous aider (bien sûr aucun ne correspond exactement), et qu'on m'a signalés ; en fait vos retours seront appréciés (et signalés ensuite).

Cours et exercices

F. Cottet-Emard, "Probabilités et tests d'hypothèses". Ed. De Boeck.

P. Bogaert "Probabilités pour scientifiques et ingénieurs". Ed. De Boeck.

S. Ross "Initiation aux probabilités", Presses Polytechniques et Universitaires Romandes

Exercices avec rappels de cours

J-M et R Morvan, "Bien débuter en probabilités", éditions Cepadues

Christophe Chesneau, "Probabilités et variables aléatoires discrètes. 430 exercices corrigés avec rappels de cours." Ed. Ellipses.

Pour s'exercer sur la partie du cours "Ensembles et événements"

Chapitre 1 du livre "Théorie des probabilités, problèmes et solutions", Presses de l'Université du Québec, de C. Reischer et al.

Et encore : Manuels de cours

"Les probabilités et la statistique de A à Z", F. Dress, DUNOD

"Cours de probabilités et de statistiques", C. Leboeuf, J.-L. Roqué, J. Guégand, ELLIPSES

"Aide-mémoire de probabilités et statistiques", J. Marceil, ELLIPSES

"Probabilités", N.Boccara, ELLIPSES

"Probabilités et statistiques", A. Combrouze, PUF

Manuels contenant des exercices corrigés

"Exercices de probabilités", J. Guégand, M.-A. Maingueneau, ELLIPSES

"Exercices ordinaires de probabilités",G. Frugier, ELLIPSES

"Probabilités. Exercices corrigés", D. Ghobanzadeh, TECHNIP

"Calcul des probabilités : Cours, exercices et problèmes corrigés", D. Foata, A. Fuchs, DUNOD

1 Introduction et motivations

Le but officiel de ce cours est de vous apprendre un peu de statistiques. L'utilité d'avoir quelques notions en statistique n'est plus à démontrer, puisqu'il y en a partout et en particulier en sciences, et souvent assez mal digérée on dirait.

Voici une question simple à laquelle vous saurez répondre à la fin du cours. Désolé de prendre cet exemple. On veut évaluer le pourcentage de personnes atteintes d'un certain virus dans une population de 50 000 000 de personnes. On dispose d'un test complètement fiable, qu'on fait sur 1000 personnes totalement tirées au hasard et représentatives (évidemment, c'est la partie la plus dure à faire : pas question de choisir seulement les étudants de la fac ou les personnels d'un club de foot !). On trouve 5% de test positifs. Quelle est la probabilité qu'en fait plus de 10% de la population soit atteinte ?

Bon, on ne fera pas que cela. On aura besoin d'un certain nombre de notions qui en fait on un intérêt en soi, on apprendra à modéliser des situations simples, et bien entendu on ne peut pas faire de statistique sans faire un peu de probabilités avant (et en fait ça tombe plutôt bien, parce que c'est intéressant en soi).

Qu'est-ce que j'entends par modéliser un phénomène aléatoire ? On veut par exemple savoir quelles sont les chances (quelle est la probabilité), quand on lance un dé 5 fois de suite, d'obtenir exactement cinq fois un 6. Le modèle le plus simple à construire est de dire que l'ensemble des réalisations possibles des six lancers est représenté par l'ensemble Ω des quintuplet $\omega = (a_1, a_2, a_3, a_4, a_5)$, où chacun des cinq a_i est un nombre entier compris entre 1 et 6. Donc Ω est un ensemble qui a 6^5 éléments ω ; si les dés sont normaux et les lancers honnêtes, chacun des 6^5 éléments a la même probabilité 6^{-5} de se produire, et un moyen de calculer la probabilité ci-dessus est ce compter pour combien de ω la propriété correspondante est réalisée. Donc ici il y a 25 possibilités (5 choix de numéros pour le lancer malheureux, et pour chacun 5 choix de nombre entre 1 et 5. On divise par 6^5 et on obtient une probabilité 25×6^{-5} assez faible.

Mini-exercice (pour vous aider à vérifier que c'est clair) : quelle probabilité de faire au moins cinq fois un 6 ?

Mini-exercice : avec quel ensemble modéliseriez vous 14 lancers de pièces pour jouer à pile ou face ?

Evidemment pour des cas un peu plus compliqués on sera amenés à faire de raisonnements combinatoires plus compliqués aussi (donc un peu de technique ne nuira pas), mais l'idée est là. En tout cas des raisonnements pouvant aider à calculer plus vite seront utiles, et on devra aussi apprendre à ne pas les appliquer à tort.

Remarque. Ceci traduit peut-être seulement mon point de vue personnel sur les statistiques, mais je le dis quand même. Pour moi l'utilité principale est de tirer les conclusions les plus complètes possibles des informations, souvent très incomplètes, qu'on a à sa disposition. Les exemples qui suivent vont un peu dans ce sens : plus on a d'informations, plus on peut dire de choses. Mais aussi, parfois les stats aident à obtenir des informations qui seraient difficiles à deviner directement.

Expérience qu'on pourrait faire en cours. Demander à chacun d'écrire une suite aléatoire,

composée de, disons, vingt chiffres égaux à 1 ou 2. Par exemple 21211212221221121211. Puis demander à l'audience :

- combien ont commencé par 12 ou 21 ? (ca devrait être 50%)

- combien contiennent au moins une suite de 5 chiffres identiques ? J'ai pas calculé, mais ca devrait être une majorité.

Pas si facile d'en faire une crédible (sans lancer des pièces). Prendre une liste ou un algorithme connu marche, sauf si je sais quelle liste ou quel algorithme....

Exemple Je vous dis qu'en Serbomalgachie, un malade de Covid sur deux est vacciné. Devez-vous en déduire que :

- 1. le vaccin ne sert à rien ?

- 2. j'essaie de vous manipuler ? (réponse : pas encore)

- 3 il me faut donner d'autres informations ? (oui en effet).

J'ajoute donc que 90% des personnes sont vaccinées. Cette fois, est-ce convaincant ? Pour moi, oui. Mais vous avez le droit de penser que je cherche à vous manipuler. Mais il vous faudra sans doute trouver une explication plus convaincante que : les chiffres sont manipulés de toute façon ; tous les vaccinés sont des trouillards, donc ils ne sortent pas, donc c'est normal qu'ils soient moins malades ; la moitié des non-vaccinés ont la maladie génétique X qui les empêche de se faire vacciner, et qui les rend aussi particulièrement sensibles à la Covid. Vous remarquerez qu'en parlant de cela, on calcule implicitement plein de probabilités (comme celle que des millions de personnes participent à un complot planétaire, mais qu'aucune n'ai jamais pensé à tout raconter).

Exemple. Imaginons que le RER est organisé de la manière suivante : un train part, disons de la gare du Nord, toutes les 20 minutes, se dirige vers St Rémi, puis fait demi-tour, revient gare du nord, etc. Donc à Orsay il passe un train toutes les 20 minutes dans chaque direction. Je vais à Orsay sans regarder ma montre, et prends le premier train qui passe. Quelle est ma probablité de me retrouver à St Rémi ? Ne dites pas 1/2. Dites "je n'ai pas assez d'information" ; c'est souvent cela la bonne réponse. Par contre, si je vous dis qu'il faut 39 minutes pour un train qui passe par Orsay en direction de St Rémi pour y aller, faire demi-tour, et revenir à Orsay, là vous pouvez répondre : j'ai 1 chance sur 20 d'aller vers St Rémi (pourquoi ?). Et encore, en faisant toutes sortes d'hypothèses implicites.

L'une des erreurs les plus classiques dont on devra se méfier est de considérer que des événements (par exemple "je suis écolo", ou "j'aime la moto") sont indépendants, alors qu'ils ne le sont pas.

Une illustration maintenant assez classique pour vous amuser, issue d'un jeu télévisé. J'essaie de gagner une voiture. Elle est cachée derrière une porte parmi trois. Je fais un premier choix parmi les 3. Le présentateur ouvre alors l'une des deux autres portes (bien sûr qui n'est pas celle derrière laquelle est la voiture). Je peux alors garder mon choix initial ou choisir la porte que n'a pas ouverte le présentateur. Et si j'ai choisi la porte derrière laquelle est la voiture, j'ai gagné la voiture. Quelle est l'option qui me donne le plus de chances de gagner ? Quelle est ma probabilité de gagner ? Réponse en TD.

Exemple. On sait (c'est ce que dit mon prédécesseur, je n'ai pas retrouvé la stat) qu'un conducteur sur 1000 est alcoolisé. Et que le l'alcootest donne un pour cent de faux positifs.

On teste un conducteur au hasard et on trouve un résultat positif. Quelle est la probabilité qu'il soit réellement alcoolisé ? Pourquoi n'est-ce pas forcément aussi grave que cela semble ? [Résultat = une chance sur 10].

Exemple. Un voleur a eu la mauvaise idée de laisser traîner ses empeintes sur le lieu du crime. Ca tombe bien, j'ai une super-base de données avec six millions d'empreintes, alors je regarde, et je trouve une coincidence, avec une chance sur 10^7 que l'empreinte ne corresponde pas. Dois-je immédiatement boucler l'affaire ?

Ce cours a un parti pris : les probas discrètes. On fera presque tous nos calculs sur des ensembles finis (comme les exemples précédents). Ce n'est pas une théorie complète (il est aussi utile d'avoir des mesures sur $[0,1]$ ou sur $\mathbb{R}$), mais cela nous permettra d'avoir une théorie plus simple, et quand même de comprendre les principes essentiels.

2 Ensembles et dénombrement

A la fois des notations, définitions, et moyens de calcul qui nous seront très utiles.

2.1 Ensembles, cardinal, injections, surjections, bijections

On profite des définitions qui suivent pour rappeler aussi d'autres notations standards, donc si certaines écritures vous paraissent étranges, familiarisez vous.

Souvent on choisira en premier un grand ensemble Ω qui contient tous les autres (qu'on considère), et on appellera Ω un univers. Les éléments de Ω seront typiquement notés ω, et on parfois dira que ω est un <u>événement élémentaire</u>.

Et les ensembles (donc souvent contenus dans Ω) seront notés par une lettre majuscule, comme A ou B, et il seront parfois appelés événements. On reviendra sur ces noms bizarres quand on parlera de probas.

J'ajoute une notation personnelle qui me permettra de gagner du temps. Pour chaque entier $n \geq 1$, on note

$$E_n = \{1, 2, \ldots, n\}$$

l'ensemble des entier compris entre 1 et n. Ce sera notre ensemble à n élément le plus simple.

Exemple. Pour modéliser 3 tirages de dés standards, on utilisera par exemple l'ensemble $\Omega = (E_6)^3 = E_6 \times E_6 \times E_6 = \big\{(a_1, a_2, a_3)\,;\, a_1 \in E_6, a_2 \in E_6, \text{ et } a_3 \in E_6\big\}$. On aurait pu prendre un Ω plus grand, comme par exemple une longue suite de nombres représentant la position de toutes les particules du monde en tout temps rationnel, mais ce serait peu pratique et totalement inutile. Un exemple d'événement est $A = \big\{(a_1, a_2, a_3) \in (E_6)^3\,;\, a_1 + a_2 + a_3 = 7\big\}$, qui correspond à l'événement "j'ai fait un total de 7".

On supposera le plus souvent que <u>tous nos ensembles sont finis</u> (ont un nombre fini d'éléments). Le <u>cardinal</u> de A, noté $\mathrm{Card}(A)$ ou encore $\sharp A$ est alors le nombre d'éléments de A. Je prends le parti pris, maintenant et dans ce qui suit, de ne pas définir tout ceci de manière plus précise : vous savez bien ce que c'est. Je donne quand même sans démonstration un certain

nombre de propriétés "logiques", qui en fait se démonteraient par récurrence en faisant juste attention à l'ordre dans lequel on démontre tout. D'abord

• Si $A \subset B$ (deux ensembles finis), alors $\sharp A \leq \sharp B$, et si de plus $\sharp A = \sharp B$, alors $A = B$.

Maintenant considérons des applications $f : A \to B$. On rappelle que f est dite injective si $f(a_1) \neq f(a_2)$ pour tout choix de $a_1, a_2 \in A$ tels que $a_1 \neq a_2$. Dit de manière plus compliquée (mais très utile en pratique), ceci signfie que pour $a_1, a_2 \in A$, on a l'implication suivante : "$f(a_1) = f(a_2)$" $\Longrightarrow$ "$a_1 = a_2$".

On dit que f est surjective quand pour tout $b \in B$, il existe $a \in A$ tel que $f(a) = b$. Autrement dit, pour tout $b \in B$, l'image inverse $f^{-1}(\{b\}) = \big\{a \in A\,;\, f(a) = b\big\}$, n'est pas vide. Ou encore, l'image de f, à savoir l'ensemble $f(A) = \big\{f(a)\,;\, a \in A\big\}$ est égale à B tout entier.

On dit que f est bijective quand elle est à la fois injective et surjective. Donc quand pour tout $b \in B$, il existe un $a \in A$ unique tel que $f(a) = b$ (l'existence parce que f est surjective, l'unicité parce qu'elle est injective).

Mini-exercice : dessiner des patates, des croix, et des flèches (devrais-je dire des diagrammes ?) correspondant aux trois situations et contre-exemples.

Revenons aux cardinaux, et énoncons les propriétés faciles promises. Dans ce qui suit, A et B sont des ensembles finis.

• S'il existe une application bijective $f : A \to B$, alors $\sharp A = \sharp B$;

• En fait, $\sharp A = n$ si et seulement si il existe une bijection de E_n dans A (en bref, on peut numéroter les éléments de A !) ;

• Il existe une application injective de A dans B si et seulement $\sharp A \leq \sharp B$

• Il existe une application surjective de A dans B si et seulement $\sharp A \geq \sharp B$

• (très utile) Si $\sharp A = \sharp B$ et si $f : A \to B$ est injective, alors elle est bijective ;

• (très utile) Si $\sharp A = \sharp B$ et si $f : A \to B$ est surjective, alors elle est bijective.

Et signalons déjà que

$$\mathrm{Card}(\mathcal{P}(A)) = 2^{\mathrm{Card}(A)}, \tag{2.1}$$

où $\mathcal{P}(A)$ désigne l'ensemble des sous-ensembles de A. Démonstration plus bas.

2.2 Fonction indicatrice

Passons à la fonction indicatrice. Certains disent fonction caractéristique, mais pas les probabilistes ni les statisticiens, parce que pour eux la fonction caractéristique est autre chose (une transformée de Fourier en fait). Corrigez moi si je dis fonction caractéristique par mégarde ; en principe il n'y en aura pas dans ce cours.

Soient Ω un ensemble et $A \subset \Omega$ un sous-ensemble. La fonction indicatrice de A est la fonction notée $\mathbb{1}_A$ (ou parfois aussi χ_A), définie sur Ω, à valeurs dans $\{0, 1\}$, et définie par les conditions

$$\mathbb{1}_A(\omega) = 1 \text{ si } \omega \in A \text{ et } \mathbb{1}_A(\omega) = 0 \text{ si } \omega \in \Omega \setminus A \tag{2.2}$$

(donc $\Omega \setminus A$ est notre notation pour le complémentaire de A dans Ω ; quand il n'y a aucun doute sur qui est Ω, on pourra aussi noter $E^c = \Omega \setminus A$).

Notez que si A et B sont disjoints, alors $1_{A\cup B} = \mathbb{1}_A + \mathbb{1}_B$; c'est facile mais c'est pratique (démonstration dans une seconde). J'en profite pour rappeler que $A\cup B$ est l'union de A et B, définie avec une notation que j'utiliserai tout le temps par $A\cup B = \big\{a \in \Omega\,;\, a \in A \text{ ou } a \in B\big\}$, alors que l'intyersection est $A \cap B = \big\{a \in \Omega\,;\, a \in A \text{ et } a \in B\big\}$. Certains préfèrent mettre une barre verticale $|$, ou deux points, à la place de mon point virgule central ; chacun son gôut. Donc, pour $A, B \in \mathcal{P}(\Omega)$,

$$1_{A\cup B} = \mathbb{1}_A + \mathbb{1}_B - 1_{A\cap B}. \tag{2.3}$$

Je vais le faire pour cette fois, mais les propriétés suivantes du même type sont laissées en exercice (je prétends facile). Je rappelle d'abord que ceci veut dire que $1_{A\cup B}(\omega) = \mathbb{1}_A(\omega) + \mathbb{1}_B(\omega) - 1_{A\cap B}(\omega)$ pour tout $\omega \in \Omega$. Donc je prends $\omega \in \Omega$ et je vérifie la formule en distinguant les quatre cas possibles. Si $\omega \in A$ et $\omega \in B$, les 4 fonctions indicatrices valent 1 et le résultat est vrai. Si $\omega \in A$ et $\omega \notin B$, $1_{A\cup B}(\omega) = 1$, $\mathbb{1}_A(\omega) = 1$, $\mathbb{1}_B(\omega) = 0$ et $1_{A\cap B}(\omega) = 0$, donc la formule est vraie. Si $\omega \notin A$ et $\omega \in B$, on fait pareil (en échangeant les rôles de A et B) ; je vous laisse faire. Enfin si $\omega \notin A$ et $\omega \notin B$, toutes les fonctions indicatrices sont nulles et la formule est encore vraie. Notez au passage que (avec le même genre de démonstration ; distinguer 2 cas)

$$1_{A^c} = 1 - \mathbb{1}_A \quad \text{(on rappelle que } A^c = \Omega \setminus A\text{)}. \tag{2.4}$$

Plus généralement, on dit que les ensembles $A_1, \dots A_n$ forment une <u>partition de l'ensemble A</u> quand A est l'union disjointe des A_j. Disjointe signifie que les A_j sont deux à deux disjoints, ou autrement dit $A_i \cap A_j = \emptyset$ quand $i \neq j$. Et si $A_1, \dots A_n$ forment une partition de A, alors

$$\sum_{j=1}^{n} \mathbb{1}_{A_j} = 1_A,$$

ou dit de manière plus explicite, $\sum_{j=1}^{n} \mathbb{1}_{A_j}(x) = 1_A(x)$ pour tout $x \in \Omega$. Et en effet, les deux membres sont nuls si x n'est pas dans A, alors que si $x \in A$, le second membre contient exactement un terme non nul (égal à 1), correspondant à l'unique A_j qui contient x. Dans le cas où $A = \Omega$, on trouve que

$$\sum_{j=1}^{n} \mathbb{1}_{A_j} = 1 \quad \text{quand les } A_j \text{ forment une partition de } \Omega.$$

Parfois on utilise $\mathbb{1}_A$ pour compter les éléments de A : vérifiez que

$$\sharp A = \sum_{w\in\Omega} \mathbb{1}_A(\omega) \tag{2.5}$$

et en déduire par exemple que $\sharp(A \cup B) = \sharp A + \sharp B - \sharp(A \cap B)$.

Enfin démontrons (2.1) (le fait que $\mathrm{Card}(\mathcal{P}(\Omega)) = 2^{\mathrm{Card}(\Omega)}$) à l'aide de la fonction indicatrice. D'abord notons que la fonction indicatrice de A caractérise A, au sens où l'on peut retrouver A à partir de $\mathbb{1}_A$, puisque A est l'ensemble des $\omega \in \Omega$ tels que $\mathbb{1}_A(\omega) = 1$. De plus, pour toute function $f : \Omega \to \{0, 1\}$, il existe bien un ensemble (unique) $A \subset \Omega$ tel que $f = \mathbb{1}_A$, à savoir l'ensemble des $\omega \in \Omega$ tels que $f(\omega) = 1$. Je viens de dire laborieusement qu'il y a une bijection entre $\mathcal{P}(\Omega)$ et l'ensemble des applications de Ω dans $\{0, 1\}$. Donc il y a autant de parties de Ω que de fonctions $f : \Omega \to \{0, 1\}$, et on va voir que ce nombre est $2^{\mathrm{Card}(A)}$. En fait vérifions carrément un peu plus.

Proposition 2.1. *Si A et B sont des ensembles finis*

$$\text{il y a exactement } (\sharp B)^{\sharp A} \text{ applications } f : A \to B. \tag{2.6}$$

On fixe B et on démontre ceci par récurrence sur le nombre d'éléments de A. Si A a un seul élément a, il s'agit juste de définir $f(a)$ et il y a bien $\sharp B = (\sharp B)^1$ possibilités. Donc la formule est vraie quand $\sharp A = 1$.

Et si la formule est vraie quand $\sharp A = n$ (avec $n \geq 1$), on va montrer qu'elle est encore vraie quand $\sharp A = n + 1$. On se donne A tel que $\sharp A = n + 1$. On choisit n'importe quel élément $a \in A$. Et on observe que choisir $f : A \to B$, c'est exactement la même chose que choisir la restriction de f à $A \setminus \{a\}$ (choisir les valeurs de f sur le complémentaire de $\{a\}$), et ensuite choisir $f(a)$. Pour la restriction de f à $A \setminus \{a\}$, l'hypothèse de récurrence dit qu'on a exactement $(\sharp B)^n$ choix. Et pour chacun de ces choix, on a encore $\sharp B$ choix pour $f(a)$. En tout, cela fait $(\sharp B)^n \times \sharp B$ choix, donc bien $(\sharp B)^{n+1}$ comme annoncé. □

On finit par des applications faciles des propriétés de 1_A sur les cardinaux, à démontrer par exemple en utilisant (2.5) (on pourrait le faire directement !) :

- $\mathrm{Card}A^c = \mathrm{Card}\Omega - \mathrm{Card}A$;
- $\mathrm{Card}(A \cup B) = \mathrm{Card}A + \mathrm{Card}B - \mathrm{Card}(A \cap B)$;
- Si A_1, A_2, $\dots A_m$ forment une partition de A, alors $\mathrm{Card}A = \sum_{j=1}^{m} \mathrm{Card}A_j$.

On a déjà parlé de la première ; la seconde est encore plus simple, et pour la troisième il s'agit de calculer une grande somme en regroupant par paquets.

2.3 Ensembles produits

On en a déjà vu un ou deux (pas pu m'empêcher). Donc si A et B sont deux ensembles, le produit est

$$A \times B = \big\{(a, b)\,;\, a \in A \text{ et } b \in B\big\} \tag{2.7}$$

et c'est un ensemble qui a $\mathrm{Card}(A)\mathrm{Card}(B)$ éléments. Démonstration possible en notant que les ensembles $\{a\} \times B$, où $a \in A$, forment une partition de $A \times B$. Or chacun des ces ensembles a exactement $\mathrm{Card}(B)$ éléménts, et il y a $\mathrm{Card}(A)$ ensembles, donc c'est bon. Exercice : remplir les détails.

En général on note aussi

$$A_1 \times A_2 \dots A_N = \big\{(a_1, \dots a_n)\,;\, a_j \in A_j \text{ pour } 1 \le j \le N\big\} \tag{2.8}$$

des produits plus gros, et $(a_1, \dots a_n)$ est appelé un N-ulet (couple quand $N = 2$, triplet quand $N = 3$, etc.). On vérifie que

$$\mathrm{Card}(A_1 \times A_2 \dots A_N) = \mathrm{Card}(A_1) \dots \mathrm{Card}(A_N) = \prod_{j=1}^{N} \mathrm{Card}(A_j). \tag{2.9}$$

Ceci se démontrerait par récurrence en jouant aussi un peu avec les définitions (noter par exemple que $A \times B \times C$ peut être identifié à $A \times (B \times C)$). Pensez pour vous souvenir qu'un choix de N-uplet, c'est un choix successif de N éléments, le nombre de possibilités se multiplie à chaque fois.

Il y aurait plein de vérifications faciles à faire. Notez par exemple que si $A_1 \subset \Omega_1$ et $A_2 \subset \Omega_2$, alors la fonction indicatrice de $A_1 \times A_2 = \big\{(x_1, x_2) \in \Omega_1 \times \Omega_2\,;\, x_1 \in A_1 \text{ et } x_2 \in A_2\big\}$ dans $\Omega = \Omega_1 \times \Omega_2$ est donnée par

$$\mathbb{1}_{A_1 \times A_2}(x, y) = \mathbb{1}_{A_1}(x)\mathbb{1}_{A_2}(y) \quad \text{pour } (x, y) \in \Omega_1 \times \Omega_2. \tag{2.10}$$

Regardons quelques cas particuliers de tout ceci. Notons encore E_n notre ensemble de référence à n éléments. On s'intéresse souvent à l'ensemble E_n^p, qui est donc l'ensemble des p-uplets à valeurs dans E_n. Et par abus de langage on le notera parfois n^p ; j'essaierai d'éviter mais je ne suis pas seul au monde. Au moins, c'est cohérent, puisque le cardinal de cet ensemble produit est n^p (le produit de $n = \sharp E_n$, p fois).

Revenons encore sur (2.1) et sa généralisation (2.6). Soit E_n notre ensemble à n éléments. Une fonction de E_n dans B, c'est exactement la donnée, pour $1 \le j \le n$, d'un élement b_j de B. Donc c'est pareil qu'un n-uplet de $B^n = B \times B \dots B$ (n fois). Autrement dit, on vient de voir qu'on peut identifier l'ensemble des applications de E_n dans B avec l'ensemble B^n des n-uplets d'éléments de B. Pour moi identifier signifie trouver une bijection (naturelle) entre les deux ensembles.

Du coup, quand on compte le cardinal de ces ensembles, on trouve la même chose, à savoir $\mathrm{Card}(B)^n$ pour B^n, donc aussi pour le nombre d'applications de E_n dans B ce qui est conforme à ce qu'on a dit dans (2.6). Ici on s'est mis dans le cas où $A = E_n$, mais le cas général aurait été pareil. A cause de ceci, certains utilisent aussi la notation un peu abusive B^A pour l'ensemble des applications de A dans B (en tout cas, le nombre d'éléments correspond, et on a un moyen "canonique" d'identifier ces deux ensembles).

Exemple : à nouveau une remarque sur (2.1). On a dit que se donner une partie de Ω (un ensemble $A \subset \Omega$), c'est pareil que se donner une fonction de Ω dans $\{0, 1\}$ (ou E_2 si vous préférez, ajoutez 1 à la fonction). Donc bien sûr il y a autant de parties de Ω que d'applications $f : \Omega \to E_2$, à savoir $2^{\mathrm{Card}(A)}$.

2.4 Permutations de n éléménts

On se rapproche un peu de ce que j'appelle dénombrement.
On reprend mon ensemble $E_n = \{1, \dots, n\}$ à n élements.

Définition 2.2. *Une permutation à n élements est une application bijective de E_n dans E_n. On notera $\mathfrak{S}_n$ l'ensemble des permutations à n élements.*

Proposition 2.3. *Le cardinal de $\mathfrak{S}_n$ est $n! = n(n-1)\dots 1 = \prod_{j=1}^{n} j$.*
Plus généralement, si A et B sont deux ensembles à n éléments, le nombre de bijections de A dans B est toujours $n!$.

Par convention, on pose $1! = 1$ (c'est logique) et même $0! = 1$ (ça peut vous paraître moins logique mais ça marche mieux dans les formules ; l'idée est qu'un produit avec 0 termes dedans vaut 1, tout comme une somme avec 0 termes vaut 0).

Bien sûr, si A et B n'ont pas le même nombre d'éléments, il n'y a pas de bijections. C'est le second cas qui nous intéresse en général. Par exemple, combien y a-t-il de moyens d'asseoir 127 personnes dans une salle de concert à 127 places ? Clairement, c'est pareil que le nombre d'applications bijectives de l'ensemble des 127 personnes à asseoir vers l'ensemble des sièges de la salle. On se rend compte aussi, en numérotant les personnes et les sièges, que ca revient à compter les applications de E_n dans E_n.

Commençons la démonstration par le cardinal de $\mathfrak{S}_n$. La différence avec le nombre des applications vu plus haut, c'est qu'on veut maintenant qu'elles soient injectives. On compte donc le nombre de f injectives, de E_n dans E_n. Déjà, il y a n valeurs possibles pour $f(1)$. Et, pour chacune de ces valeurs de $f(1)$ (pensez que $f(1)$ a été choisi), il y a exactement $n-1$ valeurs possibles de $f(2)$ (on doit éviter $f(1)$). Bref, il y a $n(n-1)$ choix possibles du couple $(f(1), f(2))$.

Pour chacun de ces choix, il y a exactement $n-2$ choix possibles de $f(2)$ (on doit éviter les deux premières valeurs), ce qui fait exactement $n(n-1)(n-2)$ choix possibles du triplet $(f(1), f(2), f(3))$.

On continue ainsi (je vous épargne la récurence inverse qui serait juste une preuve de mauvaise foi), et on trouve qu'il y a $n!$ choix de tout le n-uplet $(f(1), \dots f(n))$. Comme promis.

Maintenant voyons comment passer du cas des bijections de E_n à celui des bijections de A dans B. Il y a une démonstration raisonable, faite en cours, qui consiste à numéroter les éléments de A, et remplacer les phrases comme "on choisit l'image de k parmi les $n-k+1$ points restants de E_n" par "on choisit l'image du k-ième élément de A parmi les $n-k+1$ points restants de B", et on voit que la preuve marche pareil.

Mais on peur aussi déduire le cas général du cas particulier de manière formelle. Je le fais ici désolé, c'est bien formel mais comme ça on aura fait au moins une fois ce genre de choses. On se donne donc A et B, tous les deux à n éléments, et on cherche le cardinal de l'ensemble $\mathfrak{S}$ des bijections $g : A \to B$.

On veut montrer que $\mathrm{Card}(\mathfrak{S}) = \mathrm{Card}(\mathfrak{S}_n)$ (et donc $\mathrm{Card}(\mathfrak{S}) = n!$), et il suffit de trouver une bijection $Z : \mathfrak{S} \to \mathfrak{S}_n$. Comme on se doute que c'est une histoire d'identification, on

commence par numéroter A et B. Donc on choisit une bijection $h_A : A \to E_n$ et une bijection $h_B : B \to E_n$.

Maintenant on définit Z. On se donne $f \in \mathfrak{S}$ et on veut définir $Z(f) \in \mathfrak{S}_n$. On fait un diagramme avec des patates et des flèches (désolé, je ne fais pas trop faire ceci en $Te\chi$) et on se rend compte que la seule chose raisonnable est de prendre $Z(f) = h_B \circ f \circ h_A^{-1}$. Ici on a noté $h_A^{-1} : E_n \to A$ la bijection réciproque de h_A et comme d'habitude $\circ$ correspond à la composition. On vérifie sans problème que $Z(f) : E_n \to E_n$, puis qu'elle est une bijection (comme composée de 3 bijections). Et voilà, on a déjà notre application $Z : \mathfrak{S} \to \mathfrak{S}_n$.

Reste à voir que c'est une bijection. Il s'agit, étant donnée $g \in \mathfrak{S}_n$, de montrer qu'il existe un unique $f \in \mathfrak{S}$ telle que $g = Z(f)$.

On commence par deviner qui c'est. On veut que $g = Z(f) = h_B \circ f \circ h_A^{-1}$. On a été raisonables dans nos définitions, et je vous laisse vérifier, en mettant vos doigts sur le dessin que je n'ai pas fait, que le membre de droite envoie bien E_n dans E_n. On compose à droite (c.-à-d. en "précomposant") les deux membres par la fonction h_A. On trouve que $g \circ h_A = h_B \circ f \circ h_A^{-1} \circ h_A = h_B \circ f$, où on a utilisé le fait que $h_A^{-1} \circ h_A$ est l'identité (sur A) par définition de la réciproque h_A^{-1}. On compose, cette fois à gauche (c-à-d. en "postcomposant"), tout le monde par $h_B^{-1} : E_n \to B$. On trouve que $h_B^{-1} \circ g \circ h_A = h_B^{-1} \circ h_B \circ f = f$, où on utilise cette fois le fait que $h_B^{-1} \circ h_B$ est l'identité sur B. Donc on a deviné qu'il faut absolument prendre $f = h_B^{-1} \circ g \circ h_A$.

Maintenant on définit $\widetilde{Z} : \mathfrak{S}_n \to \mathfrak{S}$ par la formule ad hoc, à savoir $\widetilde{Z}(g) = h_B^{-1} \circ g \circ h_A$ pour $g \in \mathfrak{S}_n$. Je vous laisse vérifier comme plus haut, en vous aidant du diagramme que vous avez dessiné, que $\widetilde{Z}(g)$ envoie bien A dans B, puis est bijective (donc $\widetilde{Z}(g) \in \mathfrak{S}$). Il ne nous reste plus qu'à voir que Z et $\widetilde{Z}$ sont deux applications réciproques, c.-à-d. que $\widetilde{Z} \circ Z$ et $Z \circ \widetilde{Z}$ sont toutes les deux l'application identique, sur $\mathfrak{S}$ et $\mathfrak{S}_n$ respectivement. On sait que c'est un moyen de vérifier que chacune des deux est une bijection (exercice !). C'est bien de suivre les calculs avec ses doigts, mais à ce stade on a vérifié que tout à un sens et il ne rete plus qu'à faire un peu d'algèbre. Pour $\widetilde{Z} \circ Z$ on se donne $f \in \mathfrak{S}$ et on calcule

$$\widetilde{Z} \circ Z(f) = \widetilde{Z}(Z(f)) = h_B^{-1} \circ Z(f) \circ h_A = h_B^{-1} \circ h_B \circ f \circ h_A^{-1} \circ h_A = f$$

où on a d'abord appliqué la définition de $\widetilde{Z}(g)$ avec $g = Z(f)$, puis appliqué la définition de $g = Z(f) = h_B \circ f \circ h_A^{-1}$, et enfin on a tout simplifié. L'autre calcul est pareil : pour $g \in \mathfrak{S}_n$,

$$Z \circ \widetilde{Z}(g) = Z(\widetilde{Z}(g)) = h_B \circ \widetilde{Z}(g) \circ h_A^{-1} = h_B \circ h_B^{-1} \circ g \circ h_A \circ h_A^{-1} = g$$

où on a commencé par calculer $Z(f)$ avec $f = \widetilde{Z}(g)$, etc. On a utilisé de multiples fois l'associativité de la composition (le fait que $f \circ (g \circ h) = (f \circ g) \circ h$), qui permet d'écrire simplement $f \circ g \circ h$.

C'était long, mais on a trouvé nos deux applications bijectives réciproques Z et $\widetilde{Z}$, donc $\mathrm{Card}(\mathfrak{S}) = \mathrm{Card}(\mathfrak{S}_n) = n!$. □

Moralité 1 : si on est raisonnablement sûr de ne pas se tromper, l'argument initial comme quoi c'est juste une histoire de renumérotation est quand même plus agréable. Moralité

2 : c'est parfois pénible de justifier pleinement ; malheureusement le risque de se tromper augmente quand on vérifie peu. J'essaierai de trouver un équilibre et ne mentir qu'à bon escient.

Fin du cours 1, 2021

2.5 Arrangements de p éléménts de $E_n = \{1, \dots, n\}$

Un cran de plus dans la science du dénombrement.

Exemple. pour expliquer ce qu'on veut : calculer combien de moyen de choisir le classement des 8 premiers étudiants parmi les 200 du cours.

Définition 2.4. *Un arrangement de p élements parmi n (ou dans E_n) est une application injective de E_p dans E_n. On note A_n^p le nombre de ces arrangements.*

On décide par convention de prendre $A_n^p = 0$ quand $n < p$ (logique : pas d'injection de E_p dans E_n dans ce cas, et même quand $p < 0$ (là c'est une vraie convention). On refera pareil pour C_n^p plus bas et je tenterai d'expliquer.

On aurait aussi pu, de manière équivalente, dire qu'un arrangement de p élements de E_n est un p-uplet $(x_1, \dots x_p) \in E_n^p$ dont toutes les coordonnées sont différentes. Donc, tel que $x_i \neq j$ pour $i \neq j$. En effet, comme précédemment, pour toute f injective on pose juste $(x_1, \dots x_p) = (f(1), \dots, f(p))$ et cela donne un p-uplet ; réciproquement, tout p-uplet vient de la fonction définie par $f(x_i) = x_i$, $1 \leq i \leq p$.

On prend $p \leq n$, parce que sinon il n'y a pas d'arrangement (ni de fonction injective de E_p dans E_n). Par contre c'est pratique de dire que $A_n^0 = 1$ (vérifiez que c'est cohérent avec la proposition ci-dessous, ou dire qu'il y a un seul moyen de choisir 0 éléments dans E_n, même en faisant attention de bien les prendre dans l'ordre).

Essayez de vous souvenir que le plus gros est en bas (j'ai du mal aussi).

Proposition 2.5. *Pour $0 \leq p \leq n$, on a*

$$A_n^p = n(n-1)\dots(n-p+1) = \frac{n!}{(n-p)!} \tag{2.11}$$

Plus généralement, pour tout choix d'ensembles A de cardinal p et B de cardinal n, le nombre d'applications injectives de A dans B est A_n^p.

On a déjà vu un cas particulier de ceci, celui où $p = n$ et il s'agit de compter les bijections. Du coup pour autoriser $p = 0$, posez aussi $0! = 1$. C'est assez cohérent (un produit sans terme).

La démonstration de (2.11) est la même que dans le cas des bijections, sauf qu'on arrête le calcul plus haut. Choisir une injection f de E_p dans E_n, c'est pareil que choisir $f(0) \in E_n$, puis $f(1)$ dans $E_n \setminus \{f(0)\}$, puis $f(2)$ dans $E_n \setminus \{f(0), f(1)\}$, et ainsi de suite.

Pour le premier choix on a n possibilités. Puis pour chacun de ces choix (c'est important que la suite du calcul ne dépende pas du choix déjà fait), on a $n-1$ possibilités pour $f(1)$,

puis à nouveau, pour chaque choix de $f(0)$ et $f(1)$, on a $n-2$ possibilités pour $f(2)$ (différent des deux autres). Et ainsi de suite. Comme on doit juste faire p choix, on s'arrête quand on a fait le produit des p premiers nombres, c.à.d avec $n-p+1$. Si $p=1$, c'est bien n (le premier terme) ; si $p=n$, on s'arrête à $n-n+1=1$, comme on a vu avant pour les permutations. Et bien sûr, si on essayait $p \geq n+1$, on n'aurait plus personne à choisir et c'est aussi pour cela qu'il n'y a pas d'injection de E_p dans E_n.

Il reste à voir, pour la seconde partie de l'énoncé, que le nombre calculé plus haut ne dépend pas des ensembles choisis. Evidemment vous n'avez par envie de revivre la démonstration de la proposition 2.3, donc on va dire que c'est pareil : une fois qu'on a choisi une numérotation $(h_A : A \to E_p)$ de l'ensemble A et une numérotation $(h_B : A \to E_n)$ de l'ensemble B, construire une application injective de A dans B, c'est pareil que de construire une application injective de E_p dans E_n ; on vient de voir qu'il y avait exactement $A_n^p = n(n-1)\dots(n-p+1) = \frac{n!}{(n-p)!}$ de faire ça.

2.6 Parties de p éléments dans $E_n = \{1, \dots, n\}$

Cette fois on demande combien de manières de choisir (un groupe de) 8 étudiants dans un amphi de 200. La différence est donc qu'on ne demande plus un classement de ces 8 étudiants.

Proposition 2.6. *Pour $0 \leq p \leq n$, on note C_n^p le nombre de sous-ensembles de E_n qui ont exactement p éléments. Alors*

$$C_n^p = \frac{A_n^p}{p!} = \frac{n(n-1)\dots(n-p+1)}{p!} = \frac{n!}{p!(n-p)!}. \tag{2.12}$$

Plus généralement, pour tout ensemble A de cardinal n il y a exactement C_n^p sous ensembles de A qui ont exactement p éléménts.

Comme plus haut, poser $C_n^p = 0$ quand $p > n$ (aucun moyen de choisir p éléments parmi seulement n) et même quand $p < 0$ (une convention).

A nouveau, le gros est en bas, mais pour bien ajouter à la confusion, on utilise souvent la notation à l'envers

$$\left(\begin{array}{c} n \\ p \end{array} \right) = C_n^p = \frac{n!}{p!(n-p)!} \tag{2.13}$$

Démonstration : on va utiliser un calcul qu'on appelle "lemme du berger". C'est pas compliqué a priori. Un berger veut compter ses moutons, mais ne voit que les pattes. Il sait (parce qu'il connaît ses moutons) que chaque mouton a 4 pattes. Il trouve 64 pattes, donc il en déduit qu'il y a 16 moutons. Je lui souhaite quand même bien du plaisir à compter toutes ces pattes qui bougent.

Bon, dans notre cas, à chaque arrangement f de p éléments de notre ensemble A à n éléments, on peut associer une partie $Z = f(E_n) \subset A$ de A, qui est l'image de f, et qui a exactement p éléments.

Notons carrément $\mathcal{A}$ l'ensemble des arrangements, et $\mathcal{Z}$ l'ensemble des parties de A ayant p éléments. On veut calculer $N = \mathrm{Card}(\mathcal{Z})$ (le nombre de moutons). On sait que $\mathrm{Card}(\mathcal{A}) = A_n^p$ (le nombre de pattes). Il reste à voir comment on associe à chaque patte le mouton qui la contient, puis que chaque mouton a le même nombre de pattes.

On associe donc à chaque arrangement de p éléments dans A, et on comprend par là, à chaque application $f : E_n \to A$, l'ensemble $\Psi(f) = f(E_n)$ qui est juste l'image de f (l'ensemble des points de A choisis, en oubliant l'ordre). Et on découpe $\mathcal{A}$ en fonction de ces images. Autrement dit, pour chaque partie Z à p éléments, on note $\mathcal{A}(Z)$ l'ensemble des $f \in \mathcal{A}$ tels que $f(E_n) = Z$. Ce sont des ensembles disjoints, et leur union est $\mathcal{A}$; en fait on a juste regroupé dans $\mathcal{A}$ les arrangements qui ont une même image. Du coup, par la formule sur les partitions,

$$\mathrm{Card}(\mathcal{A}) = \sum_{Z \in \mathcal{Z}} \mathrm{Card}(\mathcal{A}(Z)). \tag{2.14}$$

Mais maintenant on va montrer que chaque $\mathcal{A}(Z)$ a exactement $p!$ éléments. En effet, pour chaque Z (fixons-le pour regarder), les éléments de $\mathcal{A}(Z)$ sont exactement les applications injectives $f : E_p \to A$ dont l'image est Z, c.-à.-d. les bijections $f : E_p \to Z$. Et la proposition 2.3 dit qu'il y en a $p!$. Bref, $\mathrm{Card}(\mathcal{A}(Z)) = p!$, et on a des moutons qui ont tous $p!$ pattes. Donc (2.14) se simplifie et donne $\mathrm{Card}(\mathcal{A}) = \sum_{Z \in \mathcal{Z}} p! = p!\mathrm{Card}(\mathcal{Z})$. Nos définitions plus haut disent que $\mathrm{Card}(\mathcal{A}) = A_n^p$ et $\mathrm{Card}(\mathcal{Z}) = N$. On trouve bien que le nombre de parties cherché est $N = \frac{A_n^p}{p!}$. Les autres formules se déduisent facilement de (2.11), et finalement l'argument montre que le nombre trouvé ne dépend pas de l'ensemble particulier choisi, mais juste de son nombre d'éléments. De sorte qu'on a bien une définition cohérente de C_n^p, et la proposition. □

Il y a plein de formules astucieuses concernant les C_n^p, dont je ne connais pas la moitié, mais il est utile d'en citer quelques unes. On commence par les plus faciles pour s'aguerrir ; ne vous inquiétez pas si ca a l'air astucieux, c'est pour ça qu'on vous les donne.

Je rappelle ma convention pour les paresseux. Plutôt que de dire que C_n^p n'est défini que pour $0 \leq p \leq n$, définissons le aussi pour $p > n$ et même, si vous insistez, pour $p < 0$, en décrétant que $C_n^p = 0$ dès que $p > n$ (ou que $p < 0$). C'est pratique pour les paresseux parce qu parfois cela évite de dire dans quel intervalle est p, et de distinguer les cas. Mais on y voit quand même un peu moins clair.

Passons aux formules. D'abord, des valeurs qu'il est bien ce connaître par coeur :

$$C_n^0 = C_n^n = 1$$

en regardant la formule, ou directement parce qu'il n'y a qu'un choix d'ensemble à 0 (ou n) éléments dans E_n. A peine plus compliqué,

$$C_n^1 = C_n^{n-1} = n$$

car il y a évidemment n moyens de choisir un élément parmi n, et pareil pour en exclure un.

Notons maintenant la propriété de symétrie (par rapport à $n/2$) :

$$C_n^{n-p} = C_n^p \quad \text{pour } 0 \leq p \leq n \tag{2.15}$$

(et aussi les autres si on utilise la convention pour paresseux). Ca se voit sur la formule avec les factorielles, mais c'est aussi vrai parce que le nombre de parties de E_n à p éléments est égal au nombre de parties à $n-p$ éléments : pour tout ensemble $A \subset E_n$ qui a p éléments, le complémentaire $A^c = E_n \setminus A$ a exactement $n-p$ éléments. Dit en termes ampoulés, l'application $A \mapsto A^c$, qui est une bijection de $\mathcal{P}(E_n)$ dans lui-même, est aussi une bijection de l'ensemble des parties de E_n avec p éléments vers l'ensemble des parties de E_n avec $n-p$ éléments. Donc les cardinaux de ces deux ensembles sont égaux.

Pour la formule suivante, rappelons que le nombre total des parties d'un ensemble à n éléments est 2^n ; c'est aussi la somme, pour $0 \leq p \leq n$, du nombre des parties de cet ensemble qui ont p éléments (on est en train de faire une partition de $\mathcal{P}(E_n)$ en morceaux), et on a déjà calculé le cardinal de chaque morceau, puisque il y a C_n^p ensembles qui ont p éléments. Donc, puisqu'on a tout calculé,

$$2^n = \sum_{p=0}^{n} C_n^p. \tag{2.16}$$

Maintenant, la formule de récurrence qui permet de calculer tous les C_n^k assez vite avec le triangle de Pascal : pour tout choix de $n \geq 1$ et $p \geq 0$, on a

$$C_{n+1}^p = C_n^p + C_n^{p-1}. \tag{2.17}$$

Pour éviter de distinguer des cas, on utilise notre convention pour paresseux, qui dit que $C_n^p = 0$ quand $p = -1$. Normalement, je n'aurais pas dû écrire le coefficient C_n^{-1} correspondant au second terme quand $p = 0$, je l'ai mis quand même, mais il vaut 0. De même on a le droit de prendre $p = n+1$, mais alors le premier terme est nul.

Démonstration facile. On regarde E_{n+1}, qu'on écrit comme union de E_n et de $\{n+1\}$. Il y a deux types (disjoints) de parties de E_{n+1} qui ont p éléménts : celles qui ont leurs p éléments dans E_n, et celles qui ont $p-1$ éléments dans E_n, plus le dernier égal à $n+1$. Le premier ensemble a exactement C_n^p éléments. Le second a C_n^{p-1} (nombre de choix de $p-1$ éléments dans E_n). Noter que quand $p = 0$, le second ensemble est juste vide (et le terme correspondant est 0).

Notons aussi qu'il est assez facile de démontrer (2.16) avec les formules pleines de factorielles. C'est assez amusant (donc faites le en exercice), mais jamais vous n'auriez trouvé ceci sans le raisonnement avec les ensembles !

Je ne suis pas assez doué en $te\chi$ pour faire un beau triangle de Pascal pour calculer les premiers C_n^p ; j'aurai sans doute essayé en cours.

[Sans doute juste un bonus.] Une petite généralisation (mais qui à mon avis tombe moins souvent). On se donne A de cardinal n et B de cardinal m, disjoints, donc $A \cup B$ a $n+m$

éléments. Combien y a-t-il de parties de $A \cup B$ qui ont k éléments ? D'une part c'est C_{n+m}^k (le membre de gauche de (2.16) ci-dessous). Mais d'autre part, c'est la somme sur $p \in [0, k]$ du nombres des parties de $A \cup B$ qui ont p éléments dans A et $k-p$ éléments dans B. Pour chaque p, ce dernier nombre est le produit des deux nombres C_n^p (le nombre d'ensembles à p éléments dans A) et de C_m^{k-p} (le nombre d'ensembles à $k-p$ éléments dans B) parce que chaque couple de choix (p éléments dans A, et $k-p$ dans B) est associé un eu un seul choix global, et réciproquement. Donc on trouve que

$$C_{n+m}^k = \sum_{p=0}^{k} C_n^p C_m^{k-p}. \tag{2.18}$$

Pour gagner de la place, on a utilisé la définition élargie des C_n^p, par exemple qui est que $C_n^p = 0$ quand $p > n$. Comme ca on n'a pas besoin de distinguer les cas pour savoir s'il est possible ou non, pour chaque p, de mettre p éléments sans A et $k-p$ dans B : si on ne peut pas, le terme correspondant est 0 mais on l'écrit quand même dans la somme de (2.18).

Revenons aux formules importantes : voici la formule du binôme, qui est très pratique (surtout que le triangle de Pascal permet de recalculer les coefficients assez vite), et qui dit que pour $a, b \in \mathbb{R}$ (ou $\mathbb{C}$ si vous voulez) et $n \geq 1$,

$$(a+b)^n = \sum_{k=0}^{n} C_n^k a^k b^{n-k}. \tag{2.19}$$

Démonstration calculatoire par récurrence. C'est vrai pour $n = 1$ (parce que $C_1^0 = C_1^1 = 1$). Supposons la formule vraie pour n et démontrons la pour $n+1$. On écrit que

$$(a+b)^{n+1} = (a+b)(a+b)^n = (a+b)\sum_{k=0}^{n} C_n^k a^k b^{n-k} = a\sum_{k=0}^{n} C_n^k a^k b^{n-k} + b\sum_{k=0}^{n} C_n^k a^k b^{n-k}.$$

Le premier terme vaut

$$a\sum_{k=0}^{n} C_n^k a^k b^{n-k} = \sum_{k=0}^{n} C_n^k a^{k+1} b^{n-k} = \sum_{p=1}^{n+1} C_n^{p-1} a^p b^{n-p+1}$$

en posant $p = k+1$. On recopie le second terme en gardant $p = k$:

$$b\sum_{k=0}^{n} C_n^k a^k b^{n-k} = \sum_{k=0}^{n} C_n^k a^k b^{n-k+1} = \sum_{p=0}^{n} C_n^p a^p b^{n-p+1}$$

On peut toujours (pour la version paresseuse) ajouter le terme $p = 0$ à la première somme, et le termes $p = n+1$ à la seconde, puisque ces termes valent 0. Et du coup en sommant les deux on trouve

$$(a+b)^{n+1} = \sum_{p=0}^{n+1} [C_n^{p-1} + C_n^p] a^p b^{n-p+1}.$$

Tout ceci tombe très bien : on utilise la formule (2.17) pour remplacer $[C_n^{p-1} + C_n^p]$ par C_{n+1}^p et on trouve $(a+b)^{n+1} = \sum_{p=0}^{n+1} C_{n+1}^p a^p b^{n-p+1}$, qui est bien la formule (2.19) pour $n+1$.

Démonstration conceptuelle de (2.19). On développe $(a+b)^n$. Pour chaque terme la somme des puissances est bien n, donc on sait qu'on va obtenir $(a+b)^n = \sum_{k=0}^n c(k) a^k b^{n-k}$ en regroupant les termes, et il s'agit de calculer les coefficients $c(k)$.

Il y a 2^n termes en tout (tiens, $\mathrm{Card}(\mathcal{P}(E_n))$). Et à chacun des termes de la grande somme, on peut associer l'ensemble des numéros dans le produit où on a choisi le terme a de la somme, plutôt que le terme b. Le nombre de termes obtenus est donc le nombre d'ensembles possibles, et quand on ne regarde que les termes qui donnent $a^k b^{n-k}$, on ne regarde en fait que les ensembles qui ont k éléments (puisqu'on a pris a k fois et b le reste du temps). De sorte que $c(k) = C_N^k$.

Retour aux fonctions indicatrices Vous n'êtes pas convaincu ? Alors (bonus parce que c'est compliqué) on essaie une façon plus algébrique (mais avec des notations un peu pires) de dire tout ça. Et ceci nous amènera assez près des fonctions indicatrices.

On va donc développer $(a+b)$ de manière plus rigoureuse mais plus longue. On écrit le $j$ème du produit comme

$$a+b = a^1 b^0 + a^0 b^1 = \sum_{\varepsilon_j=0}^{1} a^{\varepsilon_j} b^{1-\varepsilon_j};$$

c'est bizarre, mais c'est vrai : distinguer deux cas. Maintenant on écrit tout le produit :

$$(a+b)^n = \prod_{j=1}^n (a+b) = \prod_{j=1}^n \sum_{\varepsilon_j=0}^{1} a^{\varepsilon_j} b^{1-\varepsilon_j}.$$

Maintenant, on développe. Il y a toujours 2^n termes, et chacun est associé à exactement une valeur du n-uplet $\underline{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots \varepsilon_n) \in \{0,1\}^n$. On trouve

$$(a+b)^n = \sum_{\underline{\varepsilon} \in \{0,1\}^n} \prod_{j=1}^n a^{\varepsilon_j} b^{1-\varepsilon_j} = \sum_{\underline{\varepsilon} \in \{0,1\}^n} a^{\sum_j \varepsilon_j} b^{n-\sum_j \varepsilon_j},$$

où pour la seconde égalité on a juste calculé l'esposant final comme somme des exposants.

Maintenant on compte les termes. On regroupe tous les termes pour lesquels on a le terme $a^k b^{n-k}$, c.-à.-d. ceux pour lesquels on a $\sum_{j=1}^n \varepsilon_j = k$. On doit montrer qu'il y en a exactement C_n^k. Donc notre problème est devenu le suivant : démontrer que

(2.20)

$$\text{le nombre de } n\text{-uplets } \underline{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots \varepsilon_n) \in \{0,1\}^n \text{ tels que } \sum_{j=1}^n \varepsilon_j = k \text{ est exactement } C_n^k.$$

On peut traduire ceci en termes de fonctions indicatrices. On a vu plus haut qu'un n-uplet $\underline{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots \varepsilon_n) \in \{0,1\}^n$, c'est pareil qu'une fonction $\varepsilon : E_n \to \{0,1\}$ (à chaque j, on

associe $f(j) = \varepsilon_j$). Donc on calcule $\sum_j f(j)$, et on demande pour combien de fonctions cette somme vaut k.

Et maintenant, on sait aussi qu'une fonction $\varepsilon : E_n \to \{0, 1\}$, c'est pareil qu'une fonction indicatrice d'ensemble (l'ensemble $A \subset E_n$ des j tels que $f(j) = 1$. Mais quand on écrit $f = \mathbb{1}_A$, la somme $\sum_j f(j)$ devient la somme $\sum_{j \in E_n} \mathbb{1}_A(j)$, qui est exactement le cardinal de A (voir (2.5) et se souvenir qu'ici $\Omega = E_n$). Donc en fait, on demande quel est le nombre de fonctions indicatrices de somme k, ce qui est pareil que le nombre d'ensembles de cardinal k. Et ce nombre est bien C_n^k par définition.

3 Espaces de probabilité

3.1 Univers et événements

On essaie toujours de modéliser des "expériences", comme des tirages à pile ou face, mais aussi des choses plus compliquées comme la situation de populations vis-à-vis de certaines données.

On commence par définir un univers Ω. C'est un ensemble, et pour nous le plus souvent il sera fini. On y pense comme à une description ou une liste de tout ce qui peut arriver qui nous intéresse.

Fin de cours 2 en 2020

Exemple 1. Si on veut faire 7 lancers de pièce, le plus simple est de considérer $\Omega = E_2^7$, qui a l'avantage d'être simple. Ca sera bien si on décide que tout ce qui nous intéresse est la liste des tirages successifs (donc une suite de 0 et de 1), et sans doute aussi qu'on ne veut pas trop savoir comment chaque pièce a été lancée et par qui, et si elle risque de rester sur la tranche, et ainsi de suite. Si on sait qu'avec de l'entraînement, on peut se débrouiller pour affecter les probabilités de pile et faces, et qu'on veut en ternir compte, on sera sans doute bien inspiré de prendre un univers plus détaillé, par exemple $\Omega = E_m^7$, où l'on a ajouté pour chaque tirage le numéro de la personne qui a lancé la pièce, un indicatif de son humeur au moment du tirage, la force du vent, etc.

Mais aussi, par ailleurs, rien n'empêche de prendre volontairement un univers plus grand, comme $\Omega = E_2^7 \times E_6^9$. On ne sait jamais ; comme ça on pourra toujours modéliser quelques lancers de dés en plus pour le même prix. Et en principe, la seconde partie du produit ne gênera pas trop dans les calculs. Ou on veut se simplfier la vie : si on veut modéliser un tirage du loto (avec 5 boules tirées parmi 49, mais je ne sais pas si ce sont les bons nombres), il serait raisonnable de prendre pour Ω l'ensemble des arrangements de 5 points dans E_{49}, mais après tout rien n'empêche de prendre $\Omega = E_{49}^5$, et au moment d'attribuer les probabilités on mettra 0 sur les événement où deux des numéros sont égaux.

Ou si on s'intéresse a une variable qui pourrait être grande (le contenu en euros du compte en banque des 10 personnes les plus riches), rien n'empêche de prendre $\Omega = \mathbb{N}^{10}$ au lieu de sa fatiguer à chercher une borne a priori. Mais avec l'inconvénient que Ω sera alors infini, donc qu'il faudrait vérifier que la partie théorique du cours marche encore dans ce cas.

Si on s'intéresse aux relations entre divers indicateurs concernant une population de 10^6 personnes, on peut bien entendu prendre $\Omega = E_{10^6}$, mais il sera probablement plus pertinent de regrouper dans Ω quelques valeurs d'indicateurs dont on pense qu'ils ont un rapport avec le problème. Comme le nombre de pilules d'hydrochloroquine prises, la puissance de la trotinette utilisée, l'age de la personne, etc.

Avec notre parti pris discret, il nous manquera sans doute des cas. Par exemple, le meilleur moyen de modéliser le temps de décomposition d'une molécule radioactive est sans doute d'utiliser $\mathbb{R}$, mais même dans ce cas, une approximation par un modèle discret donnera de bons résultats (mais en fait plus compliqués !).

Suite du vocabulaire : un événement élémentaire, c'est juste un élément de Ω. Donc dans notre exemple 1, ω représente juste un jet de 7 lancers de pièces, et il y a 2^7 événements élémentaires.

Un événement, c'est juste une partie de Ω. Donc par exemple, dans l'espace $\Omega = E_2^{27}$, un événement possible est l'ensemble des $\omega \in \Omega$ dont la somme des coordonnées set 12 (correspondant au cas où on a tiré 12 fois pile et 15 fois face, par exemple).

Pour un lancer de 6 dés, "j'ai obtenu une somme de 17", ou "je n'ai pas fait de 6", sont des événements.

3.2 Mesures de probabilité

On se donne un univers Ω non vide. Pour le moment, on suppose Ω fini, mais dénombrable (c.-à-d. tel qu'il existe une bijection de $\mathbb{N}$ sur Ω) marcherait aussi, à condition de s'autoriser à calculer des sommes de séries.

Dans ce cas, pour définir une mesure de probabilité sur Ω, on a juste besoin de choisir un nombre $p(\omega) \in [0, 1]$ pour tout $\omega \in \Omega$, avec la propriété que

$$\sum_{\omega \in \Omega} p(\omega) = 1. \tag{3.1}$$

Donc on prend tous les événements élémentaires $\omega \in \Omega$ un par un, et on leur attribue à chacun une probabilité d'arriver $p(\omega)$. On normalise en disant que la probabilité totale est 1. On appelle cette fonction p initiale un germe de probabilité.

Mais dès qu'on a p, on peut tout de suite calculer la probabilité P de n'importe quel événement $A \subset \Omega$, simplement en posant

$$P(A) = \sum_{\omega \in A} p(\omega). \tag{3.2}$$

On appellera P la mesure de probabilité associée à p. Noter les propriétés simples suivantes, qu'on utilisera tout le temps. D'abord,

$$0 \leq P(A) \leq P(\Omega) = 1 \quad \text{pour tout } A \in \mathcal{P}(\Omega)\,; \tag{3.3}$$

et d'ailleurs

$$P(A) \leq P(B) \quad \text{pour tout choix de } A, B \in \mathcal{P}(\Omega) \text{ tels que } A \subset B \tag{3.4}$$

(on somme moins de termes), et

$$P(A) = \sum_j P(A_j) \quad \text{quand } A \text{ est l'union disjointe des } A_j \in \mathcal{P}(\Omega), \tag{3.5}$$

parce que $P(A)$ est la somme de tous les $p(\omega)$, $\omega \in \cup_j A_j$, qu'on peut répartir en somme partielles sur chaque A_j, puis sommer à nouveau.

On peut retrouver la fonction p à partir de P, parce que si $P : \mathcal{P}(\Omega) \to [0,1]$ est une fonction avec les propriétés ci-dessus, la fonction p définie sur Ω par $p(\omega) = P(\{\omega\})$ est un germe de probabilité, et P est la mesure de probabilité associée à p. On peut formaliser tout ceci comme suit.

Définition 3.1. *Un espace de probabilité est un couple (Ω, P), où Ω est un ensemble non vide (fini pour l'instant, dénombrable plus tard) et P est une mesure de probabilité sur Ω. Et une mesure de probabilité sur Ω, c'est juste une fonction $P : \mathcal{P}(\Omega) \to [0,1]$ qui vérifie les propriétés (3.3), (3.4), et (3.5).*

Et donc on sait que les mesures de probabilité viennent de germes de probabilité.

Exemple le plus standard : la mesure de probabilité uniforme sur Ω. On prend un ensemble fini Ω, et on pose $p(\omega) = \frac{1}{\mathrm{Card}(\Omega)}$ pour tout $\omega \in \Omega$. Facile de voir que c'est un germe de probabilité, et donc

$$P(A) = \frac{\mathrm{Card}(A)}{\mathrm{Card}(\Omega)} \quad \text{pour tout } A \in \mathcal{P}(\Omega) \tag{3.6}$$

est une mesure de probabilité sur Ω. Vérification facile, et c'est typiquement ce qu'on prend pour des lancers de pièces ou de dés (non pipés).

Mais pour un dé mal centré, on peut prendre $\Omega = E_6$, des nombres $p_i \in [0,1]$ tels que $\sum_{i=1}^6 p_i$, et la probabilité de tomber sur i sera juste p_i (rien de trop compliqué).

Retour aux formules : pour des ensembles non disjoints (on dira aussi pour des événements qui ne sont pas incompatibles), la formule n'est pas trop compliquée :

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{3.7}$$

C'est assez facile à vérifier (pensez juste à retirer les termes $p(\omega)$ qui sont comptés deux fois). Si ceci vous fait penser à la fonction de répartition, c'est normal, car en effet on peut ré-écrire

$$P(A) = \sum_{\omega \in A} p(\omega) = \sum_{\omega \in \Omega} \mathbb{1}_A(\omega) p(\omega), \tag{3.8}$$

c'est à dire qu'au lieu de sommer seulement sur $\omega \in A$, on somme tous les termes mais on multiplie au préalable par 0 tous les termes dont on ne veut pas ! Et alors on peut utiliser le fait que $\mathbb{1}_{A\cup B} = \mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_{A\cap B}$ puis sommer sur $\omega \in \Omega$ pour démontrer (3.7). On reparlera de ceci quand on intégrera les fonctions sur Ω (à traduire par "quand on on calculera

l'espérance des variables aléatoires"). En attendant, le même genre de démonstration donne aussi la formule pour le complémentaire :

$$P(A^c) = 1 - P(A) \quad \text{pour tout } A \in \mathcal{P}(\Omega). \tag{3.9}$$

En effet, $P(A^c) + P(A) = \sum_{\omega \in A} p(\omega) + \sum_{\omega \in A^c} p(\omega) = \sum_{\omega \in \Omega} p(\omega) = 1$ par les diverses définitions.

3.3 Probabilité produit

Utilité : un moyen standard de définir de nouveaux espaces probabilisés à partir d'anciens.

Voici un exemple typique. On a un modèle pour lancer de pièce ($\Omega_1 = E_2$, muni de la mesure uniforme), et un modèle pour un jet de dé ($\Omega_1 = E_6$, muni de la mesure uniforme). On suppose en outre que les deux lancers sont indépendants, au sens où on pense vraiment que le résultat du lancer de pièce n'aura aucune incidence sur le lancer de dé, ni dans l'autre sens. Alors il est raisonable de modéliser le lancer joint par la mesure produit sur l'espace produit (qui ici sera la mesure uniforme sur un ensemble à 12 éléments).

Définition 3.2. *On se donne deux espaces de probabilité (Ω_1, P_1), et (Ω_2, P_2). On note $\Omega = \Omega_1 \times \Omega_2$ l'univers produit. La mesure produit, qu'on note souvent $P = P_1 \otimes P_2$, est l'unique mesure de probabilité sur Ω qui vérifie la condition suivante :*

$$P(A \times B) = P_1 \otimes P_2(A \times B) = P_1(A)P_2(B) \quad \textit{pour tout } A \subset \Omega_1 \textit{ et tout } B \subset \Omega_2. \tag{3.10}$$

On verra l'existence tout de suite.

On rappelle que $A \times B = \big\{(a,b)\,;\, a \in A \text{ et } b \in B\big\} \subset \Omega_1 \times \Omega_2$. Donc, l'existence pour commencer. Si on veut que (3.10) soit satisfait, en particulier en regardant le cas des ensembles $A = \{\omega_1\}$, avec $\omega_1 \in \Omega_1$ et $B = \{\omega_2\}$, avec $\omega_2 \in \Omega_2$, on trouve que forcément $P_1 \otimes P_2(\{(\omega_1, \omega_2)\}) = P_1(\{\omega_1\})P_2(\{\omega_2\})$ (car l'ensemble à un élément $\{(\omega_1, \omega_2)\}$ est bien l'ensemble produit des deux ensembles $\{\omega_1\}$ et $\{\omega_2\}$. Autrement dit, si $p_1 : \Omega_1 \to [0,1]$ est le germe de probabilité correspondant à P_1 et $p_2 : \Omega_2 \to [0,1]$ est le germe de probabilité correspondant à P_2, le germe de $P_1 \otimes P_2$ doit être p, où $p : \Omega \to [0,1]$ est donné par $p(\omega_1, \omega_2) = p_1(\omega_1)p_2(\omega_2)$.

Donc déjà, on n'a pas le choix, il faut prendre pour $P_1 \otimes P_2$ la mesure de probabilité définie à partir du germe p. On doit vérifier que $p(\omega) \in [0,1]$, et aussi que $\sum_{\omega \in \Omega} p(\omega) = 1$ pour que p soit un germe. Le fait que $p(\omega) = p_1(\omega_1)p_2(\omega_2) \in [0,1]$ est clair, puisque p_1 et p_2 sont des germes, et

$$\sum_{\omega \in \Omega} p(\omega) = \sum_{(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2} p(\omega_1, \omega_2) = \sum_{(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2} p_1(\omega_1)p_2(\omega_2);$$

jusqu'ici, je n'ai fait que recopier les définitions. Mais maintenant je peux remarquer que cette somme a $\mathrm{Card}(\Omega_1)\mathrm{Card}(\Omega_2)$ termes, qu'on peut regrouper par paquets correspondant à une

valeur de ω_1 donnée. Et ensuite calculer les sommes partielles et additioner les résultats. Ceci donne

$$\sum_{(\omega_1,\omega_2)\in\Omega_1\times\Omega_2} p_1(\omega_1)p_2(\omega_2) = \sum_{\omega_1\in\Omega_1}\Big\{\sum_{\omega_2\in\Omega_2} p_1(\omega_1)p_2(\omega_2)\Big\} \tag{3.11}$$
$$= \sum_{\omega_1\in\Omega_1}\Big\{p_1(\omega_1)\sum_{\omega_2\in\Omega_2} p_2(\omega_2)\Big\} = \sum_{\omega_1\in\Omega_1}\Big\{p_1(\omega_1)\Big\} = 1$$

en utilisant la propriété de germe pour p_2, puis pour p_1. Donc dans cette fin de calcul on a utilisé le fait que pour calculer la somme de tous les éléments d'un tableau de chiffres, on peut aussi sommer les éléments de chaque ligne (ou colonne), puis sommer les sommes partielles obtenues.

Donc p est un germe, on peut définir $P = P_1 \otimes P_2$ à partir de ce germe, et il ne reste plus qu'à démontrer qu'elle vérifie la propriété plus générale (3.10). En fait on refait le même calcul que plus haut : si $A \subset \Omega_1$ et $B \subset \Omega_2$, la formule (3.2) donne

$$P(A\times B) = \sum_{\omega\in A\times B} p(\omega) = \sum_{(\omega_1,\omega_2)\in A\times B} p(\omega) = \sum_{(\omega_1,\omega_2)\in A\times B} p_1(\omega_1)p_2(\omega_2) \tag{3.12}$$
$$= \sum_{\omega_1\in A}\Big\{\sum_{\omega_2\in B} p_1(\omega_1)p_2(\omega_2)\Big\} = \sum_{\omega_1\in A}\Big\{p_1(\omega_1)\sum_{\omega_2\in B} p_2(\omega_2)\Big\}$$
$$= \sum_{\omega_1\in\Omega_1}\Big\{p_1(\omega_1)P_2(B)\Big\} = P_2(B)\sum_{\omega_1\in\Omega_1} p_1(\omega_1) = P_2(B)P_1(A),$$

exactement ce qu'on voulait pour (3.10). Donc la mesure de probas $P_1 \otimes P_2$ existe et est donnée par la formule (3.2) à partir du germe p.

L'unicité vient du fait qu'on devait choisir p comme on l'a fait, et que le germe détermine P à cause de (3.2).

La notation standard pour p est $p_1 \otimes p_2$. Donc

$$p_1 \otimes p_2(\omega_1,\omega_2) = p_1(\omega_1)p_2(\omega_2). \tag{3.13}$$

Exemple. L'exemple le plus simple est lorsque P_1 est la probabilité uniforme sur l'ensemble Ω_1 et P_2 est la pobabilité uniforme sur l'ensemble Ω_2. Dans ce cas $P_1 \otimes P_2$ est la probabilité uniforme sur $\Omega_1 \otimes \Omega_2$. C'est facile à vérifier puisque $p_1(\omega_1) = \frac{1}{\mathrm{Card}(\Omega_1)}$, $p_2(\omega_2) = \frac{1}{\mathrm{Card}(\Omega_2)}$, et par la définition (3.13) de $p = p_1 \otimes p_2$, $p(\omega_1,\omega_2) = p_1(\omega_1)p_2(\omega_2) = \frac{1}{\mathrm{Card}(\Omega_1)\mathrm{Card}(\Omega_2)} = \frac{1}{\mathrm{Card}(\Omega)}$.

Remarque. Si on veut définir $P_1 \otimes P_2 \otimes P_3$ (sur l'univers correspondant $\Omega_1\times\Omega_2\times\Omega_3$), on peut procéder comme ci-dessus. On prend le germe p défini par $p(\omega_1,\omega_2,\omega_3) = p_1(\omega_1)p_2(\omega_2)p_3(\omega_3)$ pour $(\omega_1,\omega_2,\omega_3) \in \Omega = \Omega_1 \times \Omega_2 \times \Omega_3$, puis on pose

$$P(Z) = \sum_{\omega\in Z} p(z)$$

pour tout ensemble $Z \subset \Omega$, et la vérification que c'est une mesure telle que

$$P(A \times B \times C) = P_1(A)P_2(B)P_3(C) \quad \text{pour } A \subset \Omega_1, B \subset \Omega_2, C \subset \Omega_3$$

se fait comme avant. Evidemment, ça marche aussi avec plus de termes, et aussi on aurait pu choisir de prendre $P = (P_1 \otimes P_2) \otimes P_3$, ou $P = P_1 \otimes (P_2 \otimes P_3)$, et en même temps vérifier que les deux sont égaux.

Remarque. On va revenir sur cette histoire d'indépendance, dès le prochain paragraphe, mais disons tout de suite que si la mesure qu'on a choisie sur $\Omega = \Omega_1 \times \Omega_2$ est la mesure produit, on a la formule (3.10), et on voit ceci comme une marque d'indépendance complète entre ce qui se passe sur Ω_1 et ce qui se passe sur Ω_2. Notez que la formule $P(A \times B) = P_1(A)P_2(B)$ peut aussi s'écrire, avec des notations faciles à deviner

$$P\big(\{(w_1, \omega_2)\,;\, \omega_1 \in A \text{ et } w_2 \in B\}\big) = P\big(\{(w_1, \omega_2)\,;\, \omega_1 \in A\}\big)\ P\big(\{(w_1, \omega_2)\,;\, \omega_2 \in B\}\big),$$

ou même pour gagner de l'espace,

$$P\big(\{\omega_1 \in A \text{ et } w_2 \in B\}\big) = P\big(\{\omega_1 \in A\}\big)\ P\big(\{\omega_2 \in B\}\big), \tag{3.14}$$

qui est peut-être plus parlant.

A contrario, si on sait que ce qui se passe sur les deux parties Ω_1 et Ω_2 n'est pas indépendant, ce serait maladroit de modéliser avec une mesure produit. Par exemple, si on lance deux pièces, mais qu'on fait lancer la seconde pièce à un tricheur professionnel en lui promettant une récompense si le résultat est le même que le premier lancer, on aura probablement un tableau de probabilités qui ressemble à ceci :

$$p(0,0) = p(1,1) = \frac{1+a}{4} \text{ et } p(0,1) = p(1,0) = \frac{1-a}{4}$$

avec un $a \in (0,1)$ qui dépend de l'habileté du tricheur. Notez que la somme des 4 probas est bien 1, comme il se doit, et que quand on calcule la probabilité de trouver pile au premier lancer, on trouve bien $p(0,0)+p(0,1) = 1/2$; c'est nomal, on n'a même pas encore commencé à tricher. Et la probabilité de trouver pile au second lancer est à nouveau $p(0,0)+p(1,0)1/2$, donc on n'y voit que du bleu. C'est quand on regarde la probabilité de gagner qu'on voit une différence.

Et P ne peut être une probabilité produit parce que $P(\{(0,0)\} \neq P(\{(0,0),(0,1)\}) \times P(\{(0,0),(1,0)\})$ (ce dernier vaut $1/4$).

3.4 Evénements indépendants

On poursuit de manière plus systématique cette histoire d'indépendance. Comme annoncé au début du cours, c'est une source de pièges classique, et malheureusement la notion d'indépendance qu'on va définir ne recouvre que partiellement ce qu'on entendrait par indépendance avec le vocabulaire courant.

Définition 3.3. *On se donne un espace de probabilité* (Ω, P)*, et deux événements* $A, B \subset \Omega$*. On dit que* A *et* B *sont indépendants quand*

$$P(A \cap B) = P(A)P(B). \tag{3.15}$$

Au moins c'est facile. On a vu un cas dans (3.14) quand P est une mesure produit, que A est de la forme $A_1 \times \Omega_2$, et B de la forme $\Omega_1 \times B_1$.

Exemple typique : on a lancé 17 fois un dé. On pense bien que les lancers sont indépendants. On a calculé la probabilité d'obtenir 19 à la somme des 4 premiers lancers, et 24 à la somme des 5 lancers suivant. On en déduit sans problème la probabilité de faire les deux choses à la fois.

Contrexemple typique : on a lancé le dé 3 fois. On a calculé la probabilité p_1 de faire 10 (en tout) aux deux premiers lancer (c'est assez faible) et la probabilité p_2 de faire 9 aux deux derniers lancers (un peu plus grande). Je vous parie que la probabilité de faire 10 aux deux premiers lancer et 9 au second est plus grande que $p_1 p_2$. Ca ne saute pas aux yeux, mais croire le contraire serait naif.

Encore plus clair ; la probabilité de faire 11 aux deux premiers est un certain nombre $p_1 > 0$; la la probabilité de faire 3 aux deux premiers est positive aussi (en fait c'est p_1). Mais la probabilité de faire les deux à la fois est nulle : il faudrait que le second dé donne à la fois moins de 2 et plus de 5.

Si vous voulez être certain que A et B sont des événements indépendants au sens de la définition ci-dessus, le mieux est de vous assurer qu'ils sont indépendants au sens usuel de la causalité. Ou presque (le premier lancer de dés pourrait avoir fait chauffer une face un peu plus, ce qui la rendrait plus grande mais aussi plus légère, et clairement aurait une incidence sur le second lancer. A un moment il faudra se résoudre à négliger ce genre d'effets.

Mais les choses se corsent ici : il se peut aussi que les deux événements A et B soient indépendants non pas pour une raison intelligente, mais juste parce que le calcul tombe ainsi.

Par exemple, je crois que les deux événements "faire 7 aux deux premiers lancers" et "faire 7 aux deux derniers lancers" sont indépendants, mais ceci n'est pas dû à une histoire subtile de causalité, mais plutôt une histoire de symétrie (et d'ailleurs je n'ai pas fait le calcul). Ce qui ne nous empêchera pas d'appliquer l'indépendance (3.14) pour faire des calculs.

Exemple typique fortuit On a fait un sondage sur 100 personnes en posant 2 questions (avez vous plus de 30 ans et préférez-vous le bleu ou le rouge). On tombe sur le tableau suivant :

moins de 30 ans= 12 bleu et 28 rouge

plus de 30 ans= 18 bleu et 42 rouge

Quand on additionne en lignes ou colonnes on trouve 40% moins de 30 ans, 30% de bleu, et le produit dans l'intersection. Difficile de dire que ces nombres ont une signification profonde : c'est juste que j'ai choisi les deux pourcentages globaux, et ensuite j'ai multiplié pour remplir la case "moins de 30 ans et bleu". Notez quand même que mon tableau est exactement le même que si j'avais pris le produit des deux mesures de proba (parce que quand j'ai les

nombres 40 % de moins de 30 ans, 30 % de bleu, si je prend 12 bleu moins de 30 ans, je n'ai plus le choix pour le reste du tableau. Donc quand on regarde juste les deux événements, le tableau des 4 probabilités est un tableau de mesures produits.

Passons à 3 événements. La définition est plus compliquée.

Définition 3.4. *On se donne un espace de probabilité (Ω, P), et trois événements $A, B, D \subset \Omega$. On dit que A, B, et D sont indépendants quand*

$$\begin{aligned} P(A \cap B \cap D) &= P(A)P(B)P(D); \\ P(A \cap B) &= P(A)P(B) \\ P(A \cap D) &= P(A)P(D) \\ P(B \cap D) &= P(B)P(D) \end{aligned} \tag{3.16}$$

Bref on demande l'indépendance deux à deux, et en plus une condition sur les 3.

Remarque. Les trois dernières conditions n'impliquent pas la première. Par exemple, on lance 2 dés (probabilité usuelle uniforme sur $E_6 \times E_6$. On considère les 3 événements : $A =$ le premier lancer est pair ; $B =$ le second lancer est pair, et $D =$ les deux lancers ont des parités différentes. Clairement $0 = P(A \cap B \cap D) \neq P(A)P(B)P(D)$. Pourtant, A et B sont clairement indépendants ; pour A et D (ou B et D aussi), c'est facile à calculer : une fois le premier dé lancé, on a toujours une chance sur 2 d'avoir D en lançant le second. Ou faites le calcul du nombre de possibilités si vous ne me croyez pas

Remarque. En termes de A, B, et D il y a 8 ensembles qu'on peut calculer, comme par exemple $A \cap B \cap D^c$. Donc le tableau des probabilités de ces événements est un tableau à 8 entrées. Quand on connait $P(A)$, $P(B)$, et $P(D)$, et qu'on suppose l'indépendance, on a les 4 équations ci-dessus, plus une (la somme des probabilités vaut 1), donc on a en fait 5 équations, ce qui permet a priori de retrouver les 8 nombres. En fait, c'est facile de le faire explicitement. Je dis comment ça marche (bonus).

Pour $A \cap B \cap D$, facile, la première équation dit que $P(A \cap B \cap D) = P(A)P(B)P(D)$.

Pour $A \cap B \cap D^c$, vous connaissez $P(A \cap B) = P(A)P(B)$, et $P(A \cap B \cap D)$, et maintenant comme $A \cap B$ est l'union disjointe de $A \cap B \cap D$ et $A \cap B \cap D^c$, vous trouvez $P(A \cap B \cap D^c) = P(A \cap B) - P(A \cap B \cap D)$, que vous savez calculer. Pareil pour les deux autres ensembles $A^c \cap B \cap D$ et $A \cap B^c \cap D$.

Pour $A \cap B^c \cap D^c$, vous écrivez que $A \cap B^c$ est l'union disjointe de $A \cap B^c \cap D^c$ et $A \cap B^c \cap D$. Donc $P(A \cap B^c \cap D^c) = P(A \cap B^c) - P(A \cap B^c \cap D)$. Vous venez de calculer le second, il reste à calculer $P(A \cap B^c)$. Mais A est l'union disjointe de $A \cap B$ et $A \cap B^c$, donc $P(A) = P(A \cap B^c) + P(A \cap B)$, ce qui permet de calculer $P(A \cap B^c) = P(A) - P(A \cap B) = P(A) - P(A)P(B)$. Donc vous savez calculer $P(A \cap B^c \cap D^c)$, et les deux autres avec deux complémentaires.

Il reste $P(A^c \cap B^c \cap D^c)$, mais comme c'est exactement la probabilité du complémentaire de l'union disjointe des 7 autres, c'est $1 - \Sigma$, où Σ est la somme des précédents.

Exemple. Si $\Omega = \Omega_1 \times \Omega_2 \times \Omega_3$, muni de la probabilité produit $P = P_1 \otimes P_2 \otimes P_3$, et $A_j \subset \Omega_j$ pour tout j, alors les événements

$$\widetilde{A}_1 = A_1 \times \Omega_2 \times \Omega_3 = \big\{(\omega_1, \omega_2, \omega_3\,;\, \omega_1 \in A_1\big\},$$

$$\widetilde{A}_2 = \Omega_1 \times A_2 \times \Omega_3 = \big\{(\omega_1, \omega_2, \omega_3\,;\, \omega_2 \in A_2\big\},$$

et

$$\widetilde{A}_3 = \Omega_1 \times \Omega_2 \times A_3 = \big\{(\omega_1, \omega_2, \omega_3\,;\, \omega_3 \in A_3\big\}$$

sont indépendants. Vérification assez facile à partir des définitions.

Contrexemple. Prendre $A \subset \Omega$, avec $0 < P(A) < 1$, et prendre $B = A$. On se doute que B est aussi peu indépendant de A que possible, et en effet $P(A \cap B) = P(A) \neq P(A)^2 = P(A)P(B)$. Ou alors prendre $B = \Omega \setminus A$, de sorte que $0 = P(A \cap B) \neq P(A)P(B) = P(A)(1 - P(A)$.

Remarque. En fait la seconde remarque dit que quand A, B, et D sont indépendants, les probabilités des événements comme $A \cap B^c \cap D$ se calculent exactement comme si on avait la probabilité produit. Je veux dire, si les 3 événements sont indépendants, et que vous calculez toutes les probabilités à partir de $P(A), P(B), P(C)$, vous devez trouver pareil que si vous aviez la probabilité produit. Soit parce que la remarque vaut aussi pour une probabilité produit, soit parce que vous pouvez calculer par exemple que

$$P(A \cap B^c \cap D^c) = P(A)P(B^c)P(D^c) = P(A)(1 - P(B))(1 - P(D)) \tag{3.17}$$

(et un peu pareil pour les autres ensembles). En fait j'ai bien sûr deviné la formule facilement, puisque c'est celle que l'on a pour une probabilité produit.

On revient à des choses plus simples, avec seulement deux ensembles.

Proposition 3.5. *On se donne un espace de probabilité (Ω, P), et deux événements $A, B \subset \Omega$. Si A et B sont indépendants, alors A^c et B sont indépendants.*

C'est le même calcul que plus haut. Comme B est l'union disjointe de $A \cap B$ et $A^c \cap B$, on peut calculer que $P(B) = P(A \cap B) + P(A^c \cap B)$, puis que

$$P(A^c \cap B) = P(B) - P(A \cap B) = P(B) - P(A)P(B) = (1 - P(A))P(B) = P(A^c)P(B), \tag{3.18}$$

où on a utilisé l'indépendance pour calculer $P(A \cap B)$. On en déduit bien que A^c et B sont indépendants (la définition est que $P(A^c \cap B) = P(A^c)P(B)$). □

Bien sûr, puisque "A et B sont indépendants" est une relation symétrique, ou par la même démonstration, on voit aussi que A et B^c sont indépendants, puis que A^c et B^c sont indépendants.

La formule (3.18) ressemble à (3.17), ca n'est pas surprenant. En fait, en modifiant à peine la démonstration de la proposition (juste ave des calculs plus long), on peut montrer que

$$\text{si } A, B, D \text{ sont indépendants, alors } A^c, B, D \text{ sont indépendants,} \tag{3.19}$$

et pareil pour les autres combinaisons comme A, B^c, D^c. Pas trop surprenant, mais du coup la formule (3.17) est plus naturelle.

Et aussi,

(3.20) si A, B, D sont indépendants, alors $A \cup B$ est indépendant de D,

et pareil avec $A \cap B$. Exercice sur la remarque qui permet de tout calculer.

Remarque (la dernière !) On pour $A_1, \dots A_k \subset \Omega$, peut aussi définir "$A_1, \dots A_k$ sont indépendants." Je pense qu'on n'en aura pas besoin, mais le plus simple est peut-être de demander que

$$P(Z_1 \cap Z_2 \dots Z_k) = P(Z_1)P(Z_2) \dots P(Z_k), \tag{3.21}$$

pour tout choix de $Z_1, \dots Z_k$, où chaque Z_j est soit A_j soit A_j^c. Si cela ressemble à une défnition de mesure produit, c'est normal, c'est fait exprès. C'est juste qu'ici Ω n'est pas écrit comme un produit, et qu'on a trouvé les ensembles $A_1, \dots A_k$ comme celà en plein dans Ω. Il y a des conditions plus faibles qui impliquent (3.21), mais c'est trop compliqué pour moi.

3.5 Probabilité conditionnelle

Typiquement : sachant que vous avez eu un test d'alcoolémie positif dans un contrôle routier au hasard, quelles sont les chances que vous soyez réellement alcoolisé ? On fait abstraction ici de toutes sortes d'autres données complémentaires (comme le nombre de verres que vous venez de boire).

Ou encore, quelle est la probabilité de faire de l'eczéma sachant que vous mangez des fraises chaque semaine ? Probablement cette probabilité est un peu plus élevée que la probabilité sans information sur les fraises (ou le contraire : peut-être les gens sujets à l'eczéma se méfent-ils). En tout cas c'est une notion simple et utile, complémentaire de celle d'indépendance.

Et comme plus haut il ne faudra pas y voir des notions de causalité évidente, mais juste un moyen de calcul. Ca n'est pas parce que manger du foie bio diminue un peu la probabilité d'avoir un cancer du poumon (si c'est vrai), qu'il y a une relation de cause à effet directe, ou même indirecte très subtile. Encore moins que dans votre cas précis, passer au bio vous aidera.

Définition 3.6. *On se donne un espace de probabilité (Ω, P), et deux événements $A, B \subset \Omega$. On suppose que $P(A) > 0$ (on va diviser dans une seconde). On appelle probabilité conditionnelle de B sachant A le nombre*

$$P(B|A) = \frac{P(A \cap B)}{P(A)}. \tag{3.22}$$

La notation est standard. Si $P(A) = 0$, penser que A ne se produit pas, et on ne peut que s'attirer des ennuis à essayer de calculer $P(B\big|A)$. Traduction immédiate mais qui a du sens aussi :

$$P(A \cap B) = P(A)P(B\big|A) \tag{3.23}$$

(pour être dans $A \cap B$, il faut d'abord être dans A (avec probabilité $P(A)$) puis, une fois ceci acquis, être dans B sachant A.

Remarque importante : toujours si $P(A) > 0$, A et B sont indépendants ssi $P(B\big|A) = P(B)$. Autrement dit, savoir A ne change pas la probabilité que B se produise. A nouveau, c'est juste un calcul, ça n'a pas a priori de sens philosophique.

Fin du cours 3, 2020

Proposition 3.7. *Soit donc (Ω, P) un espace de probabilité et $A \subset \Omega$ un événement tel que $P(A) > 0$. Alors l'application $B \to P(B\big|A)$ est une mesure de probabilité sur Ω, appelée mesure de probabilité conditionnelle sachant A, et qu'on notera P_A.*

Exemple. On tire un dé (proba uniforme sur E_6). La proba conditionnelle sachant qu'on a fait un nombre pair (donc on prend $A = \{2, 4, 6\}$) est donnée par le germe p_A, où $p_A(1) = p_A(3) = p_A(5) = 0$ et $p_A(2) = p_A(2) = p_A(2) = 1/3$. En gros, savoir A nous fait progresser dans la connaissance et on peut réviser nos probas. Facile à calculer et pas contre-intuitif !

Démonstration. On va commencer par calculer le germe associé. De toute façon, ce sera bien de le connaître. La définition nous oblige à prendre

$$p_A(\omega) = P_A(\{\omega\}) = \frac{P(A \cap \{\omega\})}{P(A)} \tag{3.24}$$

pour tout $\omega \in \Omega$. Si $\omega \in A$, $A \cap \{\omega\} = \{\omega\}$ et on trouve $p_A(\omega) = \frac{p(\omega)}{P(A)}$. Sinon, $A \cap \{\omega\} = \emptyset$ et on trouve $p_A(\omega) = 0$. En concentré, $p_A(\omega) = \mathbb{1}_A(\omega)p(\omega)$.

Est-ce que p_A est un germe ? Bien sûr $p_1(\omega) \in [0, 1]$ puisque $p(\omega) = P(\{\omega\}) \leq P(A)$ quand $\omega \in A$, et $p(\omega) = 0$ sinon. Il reste à voir que $\sum_{\omega\in\Omega} p_A(\omega) = 1$, et il se trouve que

$$\sum_{\omega\in\Omega} p_A(\omega) = \sum_{\omega\in A} p_A(\omega) = \sum_{\omega\in A} \frac{p(\omega)}{P(A)} = \frac{P(A)}{P(A)} = 1$$

parce que $p_A(\omega) = 0$ quand $\omega \notin A$, puis par (3.2). Donc p_A est un germe de probabilité. Alors la formule (3.2) donne une mesure de probabilité qu'on va noter $\widetilde{P}$ pour l'instant. Il ne reste plus qu'à vérifier que $\widetilde{P} = P_A$. On se donne $B \subset \Omega$, et on observe simplement que

$$\widetilde{P}(B) = \sum_{\omega\in B} p_A(\omega) = \sum_{\omega\in A\cap B} p_A(\omega) = \sum_{\omega\in A\cap B} \frac{p(\omega)}{P(A)} = \frac{P(A \cap B)}{P(A)} = P(B\big|A)$$

par (3.2) pour B, puis le fait que $p_A(\omega) = 0$ quand $\omega \notin A$, puis les diverses définitions. C'est ce qu'on voulait □

Désolé : c'est rarement des démonstrations subtiles, mais plus souvent un jeu de définition et des vérifications que tout est cohérent. Il faut "juste" éviter de se perdre dans les définitions et notations. Deux conséquences faciles :

$$P(B^c\big|A) = 1 - P(B\big|A) \tag{3.25}$$

quand $P(A) > 0$ et pour tout $B \subset \Omega$, puisque ceci peut aussi s'écrire $P_A(B^c) = 1 - P_A(B)$, et pareillement, si $P(A) > 0$ et les B_j, $1 \leq j \leq k$ forment une partition de B,

$$P(B\big|A) = P_A(B) = \sum_{j=1}^{k} P_A(B_j) = \sum_{j=1}^{k} P(B_j\big|A). \tag{3.26}$$

C'est donc bon de savoir que P_A est une mesure de probabilité. Ainsi, dans une population qui contient des ingénieurs, des boulangers, et d'autres choses, et aussi des fumeurs et des non fumeurs, on sait que la proba conditionnelle, sachant qu'on est fumeur, d'être ingénieur ou boulanger, est bien la somme des deux probas conditionnelles, sachant qu'on est fumeur, correspondant à ces deux catégories. On a décidé qu'il n'y a pas d'ingénieur boulanger.

3.6 Formule de Bayes

Les calculs seront faciles, ce sera l'interpétation qui sera un peu délicate ! Mais si vous êtes perdu (moi aussi parfois), revenez à la formule $P(A \cap B) = P(A)P(B|A)$, parce qu'on n'utilise pas grand chose de plus. Première version.

Théorème 3.8. *Soit donc (Ω, P) un espace de probabilité et $A, B \subset \Omega$ deux événements tels que $P(A) > 0$ et $P(B) > 0$. Alors*

$$P(A\big|B) = P(B\big|A)\,\frac{P(A)}{P(B)}. \tag{3.27}$$

C'est bien par $\frac{P(A)}{P(B}$ qu'il faut multiplier : dans le cas où A et B sont indépendants, on doit trouver $P(A\big|B) = P(A)$ et $P(B\big|A) = P(B)$.

De Thomas Bayes, britanique né à Londres, dans les années 1700.

Démonstration facile : $P(B\big|A) = \frac{P(A\cap B)}{P(A)}$, donc $P(B\big|A)\frac{P(A)}{P(B} = \frac{P(A\cap B)}{P(A)}\frac{P(A)}{P(B)} = \frac{P(A\cap B)}{P(B)} = P(A\big|B)$. Naturellement a supposé les dénominateurs non nuls. □

Exemple. Qu'est-ce que cela veut dire ? Prenons un exemple le plus simple possible. On a dans une population 60% de femmes et 40% d'hommes. On sait qu'il y a 30% de fumeurs dans la population générale, mais que la proportion passe à 33% chez les hommes.

On prend un fumeur au hasard, et on demande quelle est sa probabilité d'être un homme. On s'attend à un peu plus de 40%, mais combien au juste ?

On note H l'ensemble des hommes (de la population), F l'ensemble des fumeurs. Les données sont que $P(H) = 0,4$, $P(F) = 0,3$, et $P_H(F) = 0,35$. Le théorème dit, avec des

notations j'espère claires, que $P_F(H) = P_H(F)\frac{P(H)}{P(F)} = 0,33 \times (4/3) = 0,44$; donc 44% d'hommes chez les fumeurs. Statistiques inventées, naturellement, mais je me suis couvert en disant une population.

Une remarque : on a 4 inconnues (les 4 probabilités d'intersections) et 3 probas significatives, plus la règle que la somme des probas vaut 1. On doit donc pouvoir s'en sortir. Mais parfois on n'a pas directement les infos comme ci-dessus, mais par exemple à la place de la proportion totale de fumeurs, une autre donnée. D'où le second énoncé de la formule de Bayes

Théorème 3.9. *Soit (Ω, P) un espace de probabilité et soient $A, B \subset \Omega$ deux événements tels que $0 < P(A) < 1$ et $P(B) > 0$. Alors*

$$P(A\big|B) = \frac{P(B\big|A)\,P(A)}{P(B\big|A)\,P(A) + P(B\big|A^c)\,P(A^c)}. \tag{3.28}$$

Démonstration facile aussi. On écrit que $B = (B \cap A) \cup (B \cap A^c)$ (union disjointe), donc

$$P(B) = P(B \cap A) + P(B \cap A)^c = P(A)P(B\big|A) + P(A^c)P(B\big|A^c)$$

On reconnait le dénominateur de (3.26). On utilise maintenant le théorème pour conclure, ou on commence à prendre l'habitude de voir que $P(B\big|A)\,P(A) = P(A \cap B)$, puis appliquer la définition de $P(A\big|B) = P(A \cap B)/P(B)$, suivant ce qu'on préfère.

Exemple. Reprenons l'exemple des tests aléatoires d'alcoolémie. On a N conducteurs (N grand divisible par 1000). Parmi eux, 1 pour mille sont alcoolisés. On fait passer un test à chacun, avec des tests qui ont 1 chance sur 100 de donner un faux positif et 1 chance sur 1000 de donner un faux négatif.

On demande la probabilité, quand on a un test positif, d'être réellement alcoolisé. Attention, ici je ne dis pas comment on calculerait ou estimerait ces probabilités en pratique ; je me contente de dire ce qu'on sait dans notre exemple.

On note Ω notre lot de conducteurs, A l'ensemble des conducteurs alcoolisés, T celui des gens dont le test est positif. Et on met sur Ω la probabilité uniforme. Les hypothèses sont que $P(A) = 10^{-3}$, $P(T^c|A) = 10^{-3}$, et $P(T|A^c) = 10^{-2}$.

Appliquer le premier théorème serait un peu plus désagréable, parce qu'on ne connait pas $P(T)$ directement. Bon on peut se douter que $P(T)$ est de l'ordre de 10^{-2}, parce que la grande majorité de la population est sobre, mais disons qu'on veut un calcul précis et justifié.

On pourrait bien sûr faire le calcul en résolvant quatre équations, mais essayons la seconde formule à la place.

On demande $p = P(A\big|T)$. On utilise la formule et on trouve

$$p = \frac{P(T\big|A)\,P(A)}{P(T\big|A)\,P(A) + P(T\big|A^c)\,P(A^c)}.$$

On a ce qui manque assez facilement en passant aux complémentaires : $P(T|A) = 1 - P(T^c|A) = 1 - 10^{-3}$ et $P(A^c) = 1 - P(A) = 1 - 10^{-3}$. Donc (sauf erreur de calcul probable de ma part)

$$p = \frac{(1-10^{-3})10^{-3}}{(1-10^{-3})10^{-3} + 10^{-2}(1-10^{-3})} = \frac{1-10^{-3}}{1-10^{-3}+10(1-10^{-3})} = \frac{1}{11}.$$

A nouveau, si vous avez un contrôle positif et que vous ne pensez pas avoir bu, expliquez tout ceci au gendarme s'il est outré que vous demandiez une prise de sang. J'ai pris exprès des constantes pour un test plus répressif (on est presque sûr de se faire prendre, mais on risque un peu plus d'être ennuyé pour rien).

3.7 Indépendance conditionnelle

Définition 3.10. *Soit (Ω, P) un espace de probabilité et soient $A, B, C \subset \Omega$ trois événements. On suppose que $P(A) > 0$ pour pouvoir définir les probas conditionnelles. On dit que B et C sont indépendants conditionnellement à A quand*

$$P(B \cap C \big| A) = P(B \big| A) P(C \big| A). \tag{3.29}$$

Autrement dit, $P_A(B \cap C) = P_A(B)P_A(C)$. Autrement dit, B et C sont indépendants pour P_A. Définition sans surprise donc.

Exemple. On fait deux tests virologiques (indépendants et sumultanés) sur la chaque personne d'une population. Bien entendu, pour chaque personne, les résultats ne sont pas indépendants, puisque probablement le résultat sera correct. Mais on doit quand même pouvoir traduire le fait que les deux tests sont indépendants.

On s'intéresse aux faux négatifs (on ferait la même genre de calculs pour les faux positifs). On note Ω notre population, $I \subset \Omega$ l'ensemble des personnes infectées, T_1 (respectivement T_2) l'ensemble des personnes avec un premier (respectivement le second) test positif. On suppose que la probabilité de faux négatif pour un test est de 1%. Quelle est la probabilité de faux négatif pour l'ensemble des deux tests (la personne est infectée mais on a raté le test deux fois). Facile à deviner, mais on fait les calculs quand même.

Notre hypothèse est que $P(T_1^c|I) = 10^{-2}$, et que $P(T_2^c|I) = 10^{-2}$. Notre hypothèse d'indépendance est que $P(T_1^c \cap T_2^c|I) = P(T_1^c|I)P(T_2^c|I)$. On en déduit que $P(T_1^c \cap T_2^c|I) = 10^{-4}$. C'était très logique que les probabilités se multiplient, parce que l'indépendance conditionnelle est bien la bonne notion pour modéliser ce à quoi on pensait. Je crois que je veux dire que si on pense à un individu donné (et qu'on oublie un peu notre ensemble fini Ω, et notre modèle usuel où la seule manière de calculer une probabilité est de diviser le nombre d'éléments d'une partie de Ω par CardΩ), l'idée qu'on a de deux tests indépendants est exactement que le probabilité d'avoir le résultat r_1, puis le résultat r_2 soit exactement le produit $p_1(r_1)p_2(r_2)$ des deux probabilités corespondantes. Evidemment les deux nombres $p_1(r_1)$ et $p_2(r_2)$ dépendent de la personne considérée, et en particulier de leur état infectieux, donc c'est bien des probabilités conditionnelles qu'il nous faut.

Si l'on regarde les faux positifs, le même calcul nous dit que la probabilité d'avoir un double faux positif est aussi le carré de la probabilité de faux positif d'un seul test.

Il nous reste à déterminer ce que l'on fera dans le cas où les deux tests ne sont pas d'accord (envrion 2% de la population si la probabilité de faux positifs est aussi 1%).

Remarque. Ne pas penser que si B et C sont indépendants sachant A, alors ils sont indépendants sachant A^c. En gros, ce qui se passe sur A peut n'avoir aucun rapport avec ce qui se passe sur A^c. Et prendre P_A consiste à ne regarder ce qui se passe sur A (et normaliser), alors que P_{A^c} ne regarde que ce qui se passe sur A^c.

Exemple. Curieusement, il m'a fallu du temps pour en trouver un convaincant. Mais le principe est simple. Sur A, choisir deux événements indépendants (deux lancers de pièces). Sur A^c choisir deux événements non indépendants (par exemple, le même lancer de pièce). Donc par exemple lancer une première pièce disant si on est dans A ou pas, puis faire l'expérience avec deux lancers (ou un) décrite ci-dessus.

Donc on regarde la mesure de proba uniforme sur $E_2 \times E_2 \times E_2$. L'ensemble A est $\{0\} \times E_2 \times E_2$, puis $B \cap A = \{0\} \times \{0\} \times E_2$, $B \cap A^c = \{1\} \times \{0\} \times E_2$ (et donc B l'union des deux). Pour C on prend $C \cap A = \{0\} \times E_2 \times \{0\}$ (on lance la seconde pièce) mais $C \cap A^c = \{0\} \times \{0\} \times E_2$ (pas la peine, on a juste repris $C \cap A^c = B \cap A^c$). Facile maintenant de voir que sachant A, B et C sont indépendants (une chance sur deux pour chacun, une chance sur 4 pour l'intersection), mais sachant A^c, B et C sont aussi peu indépendants que possible (chacun a une probabilité $1/2$, et l'intersection aussi).

Voici une version qui a l'air à peine plus honnête, mais qui au fond revient au même. L'Égalistan est une démocratie populaire de 1000 habitants, 500 de chaque sexe, et avec deux chefs d'entreprise, un homme et une femme. Donc en Égalistan le sexe et le fait d'être chef d'entreprise sont indépendants. En déduire que c'est encore vrai dans le reste du monde serait très optimiste.

4 Variables aléatoires

4.1 Définitions

D'abord, désolé de dire ceci au risque de passer encore pour un non-probabiliste. **Une variable aléatoire, c'est juste une fonction définie sur notre ensemble** Ω. La plupart du temps, on les prendra toutes à valeurs réelles, mais dans certains cas on prend aussi $X : \Omega \to \mathbb{R}^n$ (ce qui revient à prendre plusieurs v.a. en même temps), voire autre chose.

Par exemple, prenons notre espace $E_6 \times E_6$ correspondant à deux lancers de dés. Un exemple de variable aléatoire est la fonction qui à $\omega \in \Omega$, associe la somme des deux coordonnées, donc $X : \Omega \to \mathbb{R}$ est la fonction qui à $\omega = (\omega_1, \omega_2)$ associe $X(\omega) = \omega_1 + \omega_2$. La notation X (et à défaut Y) est le plus souvent utilisée (plutôt que f comme vous auriez attendu).

Ou dans une population, une variable aléatoire peut être la richesse en dollars de chaque individu. On dit aléatoire parce que X dépend du paramètre ω, qui est considéré comme

aléatoire.

Ou dans l'ensemble Ω des élèves de la promotion, $X(\omega)$ peut être la note de ω au premier partiel. Vous voyez qu'on ne manque pas d'exemples instructifs.

Pour l'instant, pas la peine d'avoir une probabilité sur Ω, mais ca viendra vite.

Si on veut une fonctions à valeurs dans $\mathbb{R}^n$, on peut, en posant $X = (X_1, \dots X_n)$, où chaque X_j est une v.a. à valeurs dans $\mathbb{R}$. Rien n'empêcherait aussi de considérer X à valeurs dans $\mathbb{C}$, mais si on ne dit rien, ce sera $\mathbb{R}$. Et on préfère des valeurs numériques parce qu'en fait on veut les additionner et les multiplier, mais dans la définition (et la définition de la loi) ce n'est pas nécessaire.

Quelques notations. Soit $X : \Omega \to \mathbb{R}$ une v.a. On utilisera les images réciproques par X des éléments $x \in \mathbb{R}$, à savoir les ensembles

$$X^{-1}(x) = X^{-1}(\{x\}) = \big\{\omega \in \Omega\,;\, X(\omega) = x\big\} \subset \Omega, \tag{4.1}$$

où la première est un abus de notation (mais ça arrive). Mais attention, X^{-1} n'est pas une fonction : parfois $X^{-1}(x) = \emptyset$, parfois il a plein d'éléments. On aura besoin de savoir que

$$\text{les ensembles } X^{-1}(x),\ x \in X(\Omega), \text{ forment une partition de } \Omega. \tag{4.2}$$

En effet l'union est disjointe (on dit aussi que les événements sont incompatibles), parce que si $\omega \in X^{-1}(x) \cap X^{-1}(y)$, alors $X(\omega) = x$ et $X(\omega) = y$, ce qui est impossible si $x \neq y$. Et l'union de ces ensembles est Ω parce que tout ω est dans $X^{-1}(x)$ pour $x = X(\omega)$.

Notons encore que $X^{-1}(x)$ est bien défini aussi quand $x \notin X(\Omega)$; c'est juste que dans ce cas $X^{-1}(x) = \emptyset$.

On utilise aussi l'image réciproque d'un ensemble $Z \subset \mathbb{R}^n$, définie par

$$X^{-1}(Z) = \big\{\omega \in \Omega\,;\, X(\omega) \in Z\big\} \subset \Omega. \tag{4.3}$$

C'est donc une partie de Ω.

Et pour la même raison que plus haut, $X^{-1}(Z) \cap X^{-1}(Z') = \emptyset$ quand $Z \cap Z' = \emptyset$, et aussi les $X^{-1}(Z_j)$ forment une partition de Ω quand les Z_j forment une partition de $X(\Omega)$.

Il est temps de mettre une probabilité P sur Ω. On s'intéresse à la probabilité de toutes ces images réciproques, et utilise des notations parfois très peu explicites, mais que je donne quand même parce que même si je ne les utilise pas toutes, d'autres le feront. La probabilité de $X^{-1}(\{x\})$, où $x \in \mathbb{R}$ sera par exemple notée

$$P(X^{-1}(x)) = P\big(\big\{\omega \in \Omega\,;\, X(\omega) = x\big\}\big) = P\big(\{X(\omega) = x\}\big) = P\big(\{X = x\}\big) = P(X = x). \tag{4.4}$$

La dernière est franchement allusive, mais on se dit que s'il y a $P()$ autour, ce qui est à l'intérieur ne peut être qu'un sous-ensemble de Ω. Les probabilistes adorent oublier de mentionner l'argument ω, sans doute plus par souci de gagner du temps que par peur d'écrire l'aléatoire.

Et pour les images réciproques d'ensembles, on fait pareil : pour $Z \subset \mathbb{R}$, on note un peu indifféremment

$$P(X^{-1}(Z)) = P\big(\big\{\omega \in \Omega\,;\, X(\omega) \in Z\big\}\big) = P\big(\{X(\omega) \in Z\}\big) = P\big(\{X \in Z\}\big). \tag{4.5}$$

Exemple. Reprenons notre exemple où Ω représente la promotion, $X(\omega)$ la note de l'étudiant ω au premier partiel. Donc pour $x \in \mathbb{R}$, $X^{-1}(x)$ est l'ensemble des élèves qui ont eu la note x, et $P(\{X = x\})$ la probabilité pour un étudiant d'avoir la note x. Bien entendu, pour la plupart des x, cette probabilité est 0, parce que par exemple aucun étudiant n'a eu de note irrationnelle cette année. En fait il n'y a qu'un nombre fini de x tels que $P(\{X = x\}) \neq 0$, en tout cas si Ω est un ensemble fini, parce que $X(\Omega)$ est alors aussi un ensemble fini.

4.2 La loi d'une variable aléatoire

On va garder les notations suivantes qu'on ne répétera pas à chaque fois. D'abord Ω est un ensemble (univers), muni d'une mesure de probabilité P. On suppose que Ω est fini (des adaptations seraient possible, mais comme on dit on verra plus tard).

On se donne aussi une variable aléatoire $X : \Omega \to E$ (le plus souvent $E = \mathbb{R}$), en on note $X(\Omega) = \big\{X(\omega)\,;\,\omega \in \Omega\big\} \subset \mathbb{R}$ son image. C'est aussi un ensemble fini.

Pour chaque $x \in E$, on note $p_X(x) = P(X^{-1}(x) = P\big(\{X(\omega) = x\}\big)$ la probabilité que $X(\omega)$ soit égal à x. Comme en fait $p_X(x) = 0$ quand $x \notin X(\Omega)$, on pourra se contenter des $x \in X(\Omega)$, et se souvenir qu'autrement $p_X(x) = 0$. Voici la proposition qui va avec la définition suivante.

Proposition 4.1. *La fonction $p_X : X(\Omega) \to [0, 1]$ est un germe de probabilité sur $X(\Omega)$.*

On va le voir dans une minute. On aurait aimé dire un germe de probabilité sur E, mais on n'a défini les germes de probabilité que sur des ensembles finis et on pense à $E = \mathbb{R}$. Donc c'est pour cela qu'on se place sur $X(\Omega)$ (et de toute façon le reste de E ne doit pas compter). Voici la définition qui va avec

Définition 4.2. *La* **loi de la variable aléatoire** X *est la mesure de probabilité sur $X(\Omega)$ qui est définie à partir du germe de densité p_X. On la note P_X. Donc par (3.2)*

$$P_X(A) = \sum_{a \in A} p_X(a) = \sum_{a \in A} P(\big(\{X(\omega) = a\}\big) \tag{4.6}$$

pour tout $A \subset X(\Omega)$.

On démontre la proposition (qui est nécessaire pour définir P_X), puis on commente.

Il est clair que $p_X(x) \in (0, 1)$ parce que $p_X(x)$ est la probabilité d'un certain ensemble. Ensuite, on doit vérifier que

$$\sum_{x \in X(\Omega)} p_X(x) = 1.$$

Mais en fait les ensembles $X^{-1}(x)$, où $x \in X(\Omega)$, forment une partition de Ω, par (4.2). Donc $1 = P(\Omega) = \sum_{x \in X(\Omega)} P(X^{-1}(x)) = \sum_{x \in X(\Omega)} p_X(x)$ par la propriété (3.5) sur les partitions, puis la définition de p_X. C'est ce qu'il fallait vérifier. C'est bon, p_X est un germe et on peut définir P_X à partir de ce germe. La première égalité dans (4.6) est la formule (3.2) (qui donne la probabilité d'un ensemble connaissant le germe), et la seconde égalité est la définition de

p_X. Enfin, l'unicité de la mesure P_X est juste le fait qu'on peut retrouver P_X à partir de son germe p_X. □

Exemple. On reprend l'histoire des notes de la classe, et on suppose qu'on n'a mis que des notes entières comprises entre 0 et 20. Alors la probabilité pour un étudiant d'avoir eu une note comprise (au sens large) entre 12 et 14 est simplement

$$\begin{aligned} P(12 \leq X \leq 14) &= \sum_{x \in \{12,13,14\}} P(X = x) = P(X = 12) + P(X = 13) + P(X = 14) \\ &= P_X(\{12\}) + P_X(\{13\}) + P_X(\{14\}) = p_X(12) + p_X(13) + p_X(14). \end{aligned}$$

Vous voyez aussi pourquoi il est tentant d'utiliser des notations ramassées quand on est sûr(e) de savoir de quoi on parle.

Pourquoi on aime parler de loi ? Notez que la loi des notes dans la classe est la manière la plus ramassée et signifiante de parler de ces notes. Elle ne dépend pas du nom des élèves, ni de l'ordre dans lesquels on les a pris. L'idée est que cette classe, ou une autre classe "identique", aurait pu être décrite de plein de manières différentes, mais toujours avec la même loi.

Si vous tirez 6 dés, mais ne vous intéressez qu'à la somme X des nombres, le mieux est de vous intéresser directement à la loi de X, qui est une mesure de probabilité sur l'ensemble des entiers $x \in [6, 36]$, avec par exemple $p_X(6) = 6^{-6}$, $p_X(7) = 6^{-5}$, et après il faudrait faire des calculs apparamment trop durs pour moi avec des sommes de C_n^p.

Cette histoire d'invariance est assez importante.

4.3 Espérance d'une variable aléatoire

On continue à utiliser (Ω, P) et on se donne une variable aléatoire $X : \Omega \to \mathbb{R}$. Cette fois, $\mathbb{R}$ a de l'importance puisqu'on s'apprête à faire des additions et des multiplications.

Très difficile, puisque X est à valeurs numériques, de s'empêcher de calculer sa moyenne (on dira son espérance).

Définition 4.3. L'espérance de la variable aléatoire *X est le nombre réel*

$$E[X] = \sum_{\omega \in \Omega} p(\omega) X(\omega) \tag{4.7}$$

(où on rappelle que $p(\omega) = P(\{\omega\})$. Mais en fait $E[X]$ ne dépend que de la loi de X, et est aussi donnée par

$$E[X] = \sum_{x \in X(\Omega)} x p_X(x). \tag{4.8}$$

On vérifie ceci un peu plus bas. Comme la somme des coefficients $p(\omega)$ dans (4.7), et aussi des $p_X(x)$ dans (4.8) vaut 1, c'est bien une moyenne qu'on calcule. Donc quand X a des valeurs dans $[0, 12]$, par exemple, on est certain que $E[X] \in [0, 12]$.

Exemple. Obligé de prendre la roulette, puisque je crois que le vocabulaire vient des jeux. On joue à la roulette en posant une somme S sur un des nombres compris entre 0 et 36. Si ce numéro sort, on récupère $36S$ (et donc on a gagné $35S$), et sinon on perd sa mise. On pourrait modéliser en tenant compte du numéro qu'on joue, mais ne nous donnons pas cette peine, parce que l'aléatoire vient uniquement de la position finale de la boule. Donc on prend un Ω a 37 élément (le fameux 0), on suppose que le casino ne triche pas en plus, et il y a donc 36 éléments où on perd S, et un où on gagne $35S$. Tous avec la même probabilité 37^{-1} On appelle X notre gain, donc $-S$ la plupart du temps et $35S$ autrement. Puis on calcule notre éspérance de gain sur cette partie, à savoir

$$E[X] = \sum_{\omega\in\Omega} X(\omega) = \frac{1}{37}(1\times 35S + 36\times(-S)) = -\frac{S}{37}.$$

En espérant ne pas mêtre trompé dans les calcul, mais ne vous en faites pas, c'est strictement négatif.

Je vous laisse calculer que quelle que soit la manière dont vous aurez joué un certain nombre de parties avec des mises différentes et une mise totale Σ, votre espérance de gain sera $-\frac{\Sigma}{37}$. En fait, même si vous prévoyez de changer de stratégie en fonction des résultats précédents (y compris sans dire au casino comment vous faites, on gagne toujours à être un peu parano), ça ne changera rien au fait que votre espérance sera négative. C'est un des premiers théorèmes sur ce qu'on appelle les martingales, justement. Si vous avez décidé de ne vous arrêter de jouer que quand vous avez gagné 10, il y a une petite chance, mais strictement positive, que vous soyez obligé(e) de passer par un gain de -10^{100} pour y arriver, et souvent vous serez contraint d'arrêter dans des conditions potentiellement désagréabes avant.

Dans le cas des notes de nos étudiants, l'espérance de X (où $X(\omega)$ est la note obtenue par ω, $E[X]$ est juste la moyenne des notes.

Passons à la démonstration de (4.8), dont on déduira immédiatement que $E[X]$ ne dépend que de la loi de X. Il s'agit simplement de regrouper ensemble les éléments qui ont la même image, et de sommer par paquet. D'ailleurs pour calculer la moyenne à la main c'est sans doute le plus rapide (prévoir aussi de remplacer 4 notes de 9 et 4 notes de 11 par 8 notes de 10). On écrit que Ω est l'union disjointe des ensembles $X^{-1}(x)$, $x\in X(\Omega)$ (on a déjà fait ça plus haut), et donc

$$E[X] = \sum_{\omega\in\Omega} p(\omega)X(\omega) = \sum_{x\in X}\Big\{\sum_{\omega\in\Omega;X(\omega)=x} p(\omega)X(\omega)\Big\} = \sum_{x\in X}\Big\{\sum_{\omega\in\Omega;X(\omega)=x} p(\omega)x\Big\} \tag{4.9}$$

puisque $X(\omega)=x$ quand $X(\omega)=x$ (je sais, ça a l'air bête). Puis on sort x de la somme intérieure et on note que pour chaque $x\in X(\Omega)$,

$$\sum_{\omega\in\Omega;X(\omega)=x} p(\omega) = \sum_{\omega\in X^{-1}(x)} p(\omega) = P(X^{-1}(\omega));$$

on revient à (4.9), on remplace, et on trouve (4.8). □

Fin du cours 4, 2020

Donnons maintenant quelques propriétés élémentaires de l'espérance.

1. Si Y est une autre v.a., et λ, μ deux nombres réels, on définit une nouvelle v.a. $\lambda X + \mu Y$ par $(\lambda X + \mu Y)(\omega) = \lambda X(\omega) + \mu Y(\omega)$ pour tout $\omega \in \Omega$. On fait toujours ça avec les fonctions numériques. Et alors

$$E[\lambda X + \mu Y] = \lambda E[X] + \mu E[Y]. \tag{4.10}$$

Ca sera pareil quand on intégrera les fonctions, si ça n'est pas déjà fait. Et c'est très facile : on utilise (4.7) pour $\lambda X + \mu Y$, on remplace, puis on manipule les sommes. Je vous laisse faire.

2. Composer avec une fonction est aussi facile a priori : si $f : \mathbb{R} \to \mathbb{R}$ est une fonction, on définit une nouvelle v.a. Y par $Y(\omega) = f \circ X(\omega) = f(X(\omega))$, et la formule (4.7) donne directement

$$E[f(X)] = \sum_{\omega \in \Omega} p(\omega) f(X(\omega)). \tag{4.11}$$

On a un peu abusé des notations, en notant $f(X)$ au lieu de $f \circ X$. Mais on peut aussi essayer d'être malin, parce que à ce stade utiliser la loi est probablement plus rapide. Vérifions donc que

$$E[f(X)] = \sum_{x \in X(\Omega)} f(x) p_X(x), \tag{4.12}$$

cela pourra nous servir. On pourrait voir (4.12) comme un cas particulier de (4.8), mais de toute façon on aurait des termes à regrouper, donc le plus simple est de refaire la démonstration de (4.8), comme suit (mais cette fois je vais un peu plus vite)

$$\begin{aligned} E[f(X)] &= \sum_{\omega \in \Omega} p(\omega) f(X(\omega)) = \sum_{x \in X(\Omega)} \Big\{ \sum_{\omega \in \Omega; X(\omega) = x} p(\omega) f(X(\omega)) \Big\} \\ &= \sum_{x \in X(\Omega)} f(x) \Big\{ \sum_{\omega \in \Omega; X(\omega) = x} p(\omega) \Big\} = \sum_{x \in X(\Omega)} f(x) P(\{X(\omega) = x\}) = \sum_{x \in X(\Omega)} f(x) p_X(x) \end{aligned} \tag{4.13}$$

puisque $f(X(\omega))$ a la même valeur $f(x)$ dans toute la seconde somme, et comme annoncé dans (4.12).

Exemple avec la variance des notes d'une classe de 4 élèves, qui ont eu $8, 10, 10$, et 12. La moyenne est $E[X] = 10$ (je vous laisse calculer). La variance est par définition (on y reviendra)

$$E\big[(X - E[X])^2\big] = E\big[(X - 10)^2\big] = \frac{1}{4}\big((8-10)^2 + 0 + 0 + (12-0)^2\big) = \frac{1}{4}(4+4) = 2$$

(et l'écart type, qui est sa racine, est $\sqrt{2}$). Ici cela ne valait pas le coup de regrouper par notes, mais si on avait eu une classe de 40 élèves dont 10 ont eu 8, 20 ont eu 7, et 10 ont eu 13, on aurait trouvé que les trois valeurs de x ont des probabilités de 1/4, 1/2, et 1/4 et donc (sauf erreur à me signaler)

$$E\big[(X-E[X])^2\big] = E\big[(X-10)^2\big] = \frac{1}{4}(7-10)^2 + \frac{1}{2}(10-10)^2 + \frac{1}{4}(13-10)^2\big) = \frac{9}{4}+\frac{9}{4} = 9/2.$$

3. Inégalité standard : si $m \leq X(\omega) \leq M$ pour tout $\omega \in \Omega$, alors $m \leq E[X] \leq M$. Par exemple pour la majoration, $E[X] = \sum_{\omega\in\Omega} p(\omega)X(\omega) \leq \sum_{\omega\in\Omega} p(\omega)M = M$ puisque $\sum_{\omega\in\Omega} p(\omega) = 1$ par définition d'un germe.

Je vous laisse prouver que si X est constante égale à a, alors $E[X] = a$.

4.4 Quelques lois classiques

Il est utile de savoir reconnaître (ou deviner) quelques lois classiques prises par des v.a. à valeurs réelles. On donne un premier train de lois ; on en verra d'autres un peu plus loin.

1. La **loi uniforme sur un ensemble fini** $\{x_1, ...x_N\}$ de nombres. On a déjà vu ceci même quand l'ensemble en question n'est pas contenu dans $\mathbb{R}$. Chaque point x_j a la mesure $p_X(x_j) = 1/N$. L'espérance de X (où la loi de X est uniforme sur cet ensemble) est la moyenne des x_i. Si les x_j sont espacés régulièrement, $E[X] = \frac{1}{2}(x_1 + x_N)$ (exercice ; soit décrire $x_j = x_1 + (x_N - x_1)\frac{j-1}{N-1}$ et calculer, soit se ramener au cas où $x_1 + x_N = 0$ et regrouper les termes de manière symétrique.

2. La **loi de Bernoulli de paramètre** $p \in [0,1]$**.** Ici X ne prend que deux valeurs, 0 avec la probabilité $1-p$ et 1 avec la probabilité p. Exemple, tirage d'une pièce lestée, mais on peut en envisager beaucoup d'autres.

NB : Pas Bernouilli ; Bernoulli c'est pas une nouille. Même si je fais la faute presque systématiquement. Vérifions que

(4.14) $E[X] = p$ quand X est une v.a. dont la loi est une loi de Bernoulli de paramètre p.

Dém par la formule (4.8), in seul terme pour $x = 1$, qui vaut p. Donc on pourrait dire "loi de Bernoulli d'espérance p", et comme ça on n'a pas à se souvenir lequel de 0 ou 1 a la probabilité p, on le retrouve.

Utilisation théorique assez pratique : si $A \subset \Omega$,

(4.15) $\mathbb{1}_A$ est une variable aléatoire dont la loi est de Bernouilli, avec $p = P(A)$.

3. La **loi binomiale de paramètres** n **et** p (comme dans C_n^p).

Définition 4.4. *On se donne un entier $n \geq 0$ et un paramètre $p \in [0,1]$ (mais $p = 0$ et $p = 1$ sont moins intéressants). La loi binomiale avec les paramètres n et p est la loi d'une variable aléatoire X qui prend ses valeurs dans $\{0, 1, \dots n\}$, et dont la probabilité de prendre la valeur k, $0 \leq k \leq n$, est*

$$P(X = k) = p^k(1-p)^{n-k}C_n^k. \tag{4.16}$$

On connait déjà un cas, quand $n = 1$, et c'est la loi de Bernoulli de paramètre p. Noter en effet qu'alors X ne prend que deux valeurs, 0 et 1, que les deux C_n^k sont égaux à 1, et que les probabilités élémentaires sont bien p pour 1 et $1 - p$ pour 0.

Noter la manière étrange de dire, mais c'est fait exprès : on n'a pas besoin de connaître X, mais juste sa loi qui est une mesure de probabilité sur $[0, n] \cap \mathbb{N}$. Et cette mesure est la mesure de probabilité telle que la mesure de $\{k\}$ est $p^k(1-p)^{n-k}C_n^k$.

On va vérifier qu'une telle mesure existe, c.à.d. que $k \mapsto q(k) = p^k(1-p)^{n-k}C_n^k$ est bien un germe de probabilité sur $[0, n] \cap \mathbb{N}$. Il est clair que $q(k) \geq 0$, on va voir que $q(k) \leq 1$ dans une seconde, mais la vérification la plus importante est que

$$\sum_{k=0}^{n} q(k) = \sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k} = (p + (1-p))^n = 1^n = 1,$$

où l'égalité principale est obtenue en utilisant la formule du binôme (2.19) (avec $a = p$ et $b = 1 - p$). Maintenant il est clair que chaque $q(k)$ est inférieur à 1, puisque tous les termes sont positifs et la somme est 1.

Donc on a bien une mesure de probabilité, et la définition a un sens. Maintenant, d'où vient cette loi ? On verra bientôt que

(4.17) Si $X_1, \dots X_n$ sont n variables aléatoires indépendantes, et que chacune a pour loi la loi de Bernoulli de paramètre p, alors la loi de la somme $S = \sum_{j=1}^{n} X_j$ est la loi binomiale de paramètres n et p.

Bon, pour l'instant on on sait pas encore ce que sont des variables aléatoires indépendantes, alors je prends un exemple.

Exemple. On joue 100 fois à pile ou face, et on s'intéresse à la somme des résultats des lancers (disons 0 pour pile, 1 pour face). On veut que les lancers soient indépendants. On modélise l'expérience en prenant $\Omega = E_2^{100}$, avec la mesure uniforme.

Le j-ème lancer correspond à la coordonnée j du produit, et on considère la variable aléatoire X_j qui à $\omega \in E_2^{100}$ associe sa j-ème coordonée ω_j. C'est une variable aléatoire (elle prend les valeurs 0 et 1), et si on a choisi une pièce lestée (ou convexe), sa loi est une loi de Bernoulli de paramètre p (pensez déjà à $p = 1/2$ dans le cas des pièces standard). Ici notre hypothèse d'indépendance est traduite par le fait qu'on choisit sur Ω la mesure produit ; on fait toujours ça, mais on en reparlera plus tard.

Et maintenant la nombre total de fois où on a fait face, à savoir $S = \sum_j X_j$, est bien la variable qui nous intéresse.

On attend un peu pour démontrer (4.17) en général, mais en attendant démontrons que dans notre modèle de lancer de pièces, on a bien la formule souhaitée.

Donc on se donne $\Omega = E_2^n$ (pour n lancers), avec la probabilité produit de n fois a mesure de Bernoulli. Ceci signifie (on a vu un peu plus que le cas où $n = 2$, et en général prenez le

comme une définition) que (en notant q le germe puisque la lettre p est déjà prise)

$$P(\{(\omega_1, \dots \omega_n)\}) = q(\omega_1, \dots \omega_n) = \prod_{j=1}^{n} q_p(\omega_j), \tag{4.18}$$

où $q_p(\omega_j)$, le germe de la loi de Bernoulli calculé en ω_j, vaut p si $\omega_j = 1$, et $1-p$ sinon.

Comme on a dit, $X_j(\omega) = \omega_j$ (la j-ème coordonnée, et $S = \sum_{j=1}^{n} X_j$. On a tout fait pour que la loi de chaque X_j soit la loi de Bernoulli de paramètre p, et on veut calculer la loi de S.

Clairement S ne peut prendre que les valeurs entières $k \in [0, n]$, donc il s'agit de montrer que

$$P(S = k) = p^k (1-p)^{n-k} C_n^k. \tag{4.19}$$

On fixe donc $k \in [0, n]$ entier, et on commence par compter le nombre d'événements élémentaires $\omega \in \Omega = E_2^n$ tels que $S(\omega) = k$. C'est donc le nombre de $\omega \in \Omega$ qui ont exactement k coordonnées non nulles. Et il y en a autant que des sous-ensembles $A \subset E_n$ qui ont exactement k éléments : associer à chaque ω l'ensemble A_ω des numéros de ses coordonnées non nulles ; la réciproque de cette application $\omega \mapsto A$ consiste en gros à prendre la fonction indicatrice de A. Et ce nombre est C_n^k.

Donc l'ensemble des $\omega \in \Omega$ tels que $S(\omega) = k$ est composé de C_n^k événements élémentaires. Et en plus chacun d'entre eux a la probabilité $q(\omega) = \prod_{j=1}^{n} q_p(\omega_j) = p^k (1-p)^{n-k}$, puisque le produit a k termes où $q_p(\omega_j) = p$, et $n-k$ termes où $q_p(\omega_j) = 1-p$. On en déduit bien (4.19) en multipliant le nombre d'élément par la probabilité de chacun.

Fin de la démonstration de (4.17) dans le cas particulier de la mesure produit. On verra que le cas général est en fait à peu près équivalent. Si ceci vous rappelle la démonstration de la formule du binôme, c'est normal !

On sait calculer aisément l'espérance d'une v.a. de loi binomiale alors on le fait.

Proposition 4.5. *Si X a une loi binomiale de paramètres n et p, alors*

$$E[X] = np. \tag{4.20}$$

Calculons en utilisant (4.8) :

$$E[X] = \sum_{k=0}^{n} k P(S = k) = \sum_{k=0}^{n} k p^k (1-p)^{n-k} C_n^k.$$

Or $kC_n^k = k\frac{n!}{k!(n-k)!} = n\frac{(n-1)!}{(k-1)!(n-k)!} = nC_{n-1}^{k-1}$. Si $k = 0$, on a 0 des deux cotés par convention, ou alors, plus raisonnablement, on commence la somme plus haut avec $k = 1$. Du coup

$$E[X] = n \sum_{k=1}^{n} k p^k (1-p)^{n-k} C_{n-1}^{k-1} = np \sum_{k=1}^{n} k p^{k-1} (1-p)^{n-k} C_{n-1}^{k-1} = np \sum_{l=0}^{n-1} p^l (1-p)^{n-1-l} C_{n-1}^{k-1}$$

en sortant une puissance de p, puis en posant $l = k - 1$. La somme qui reste est le développement du binôme de $[p + (1-p)]^{n-1} = 1$. Il reste $E[X] = np$ comme annoncé.

J'espère que vous êtes impressionné, mais moi ça me fait plutôt flipper. Alors voici la démonstration du matheux pur qui ne sait pas calculer. On sait que $E[X]$ ne dépend pas de X tant que X a la même loi, ici une loi binomiale. Or on connait un X comme ceci, c'est la v.a. $S = \sum_{j=1}^{n} X_j$ qu'on a construite dans l'exemple précédent. Donc l'espérance cherchée est aussi $E[S]$. Ca tombe très bien, parce que maintenant

$$E[X] = E[S] = E[\sum_{j=1}^{n} X_j] = \sum_{j=1}^{n} E[X_j] = \sum_{j=1}^{n} p = np,$$

juste parce que l'espérance d'une somme est la somme des espérances. □

On fait maintenant une pause dans nos lois classiques pour parler d'indépendance.

4.5 Variables aléatoires indépendantes

Chose promise, chose due. Pour ce paragraphe, il sera plus pratique de pouvoir parler aussi de variables aléatoires à valeurs dans des ensembles quelconques (comme $\mathbb{R}^n$, ou autre chose).

Définition 4.6. *Soient $X_1, \dots X_n$ des variables aléatoires (de Ω dans un ensemble E). On dit que $X_1, \dots, X_n$ sont indépendantes quand*

$$P\big(\{\omega \in \Omega\,;\, X_1(\omega) = \alpha_1, \dots, \text{et } X_n(\omega) = \alpha_n\}\big) = \prod_{j=1}^{n} P\big(\{\omega \in \Omega\,;\, X_j(\omega) = \alpha_j\}\big) \tag{4.21}$$

pour tout choix d'éléments $\alpha_1, \alpha_2, \dots \alpha_n \in E$.

C'est une notion importante, du coup on va faire plein de commentaires, et aussi donner des définitions équivalentes.

On n'a pas demandé que E soit un ensemble de type particulier (mais on pense à $\mathbb{R}$ ou $\mathbb{R}^n$), ni qu'il soit fini, mais ce n'est pas grave. On exige (**??**) pour tout n-uplet $\alpha = (\alpha_1, \alpha_2, \dots \alpha_n) \in E^n$, mais en fait, si pour un j, on a que $\alpha_j \notin X_j(\Omega)$, alors le membre de gauche de (4.21) est nul (puisque $X_j = \alpha_j$ n'est jamais réalisé), et pareil pour le membre de droite. Donc la seule partie de (4.21) qui a un sens est quand $\alpha_j \in X_j(\Omega)$ pur tout j, ou si vous préferez, quand $\alpha = (\alpha_1, \alpha_2, \dots \alpha_n) \in X_1(\Omega) \times \dots \times X_n(\Omega)$. Et ce dernier ensemble est fini, puisque Ω est fini. Donc ca ne fait qu'un nombre fini de vérifications à faire.

Dans le même ordre d'idée, on aurait pu partir de l'hypothèse que $X_j : \Omega \to E_j$ pour des ensembles E_j différents les uns des autres, mais alors rien n'empêche de poser $E = \cup_j E_j$, de voir X_j comme étant à valeurs dans E, et d'utiliser la définition précédente. Autrement dit, j'ai fait un choix, et je suis content qu'il recouvre les autres choix raisonnables.

Ceci doit vous rappeler la notion d'ensembles indépendants. Pour tout choix de $\alpha = (\alpha_1, \alpha_2, \dots \alpha_n)$, notons $B_j = X_j^{-1}(\alpha_j) = \big\{\omega \in \Omega\,;\, X_j(\omega) = \alpha_j\big\}$. Alors (4.21) s'écrit aussi

$$P(B_1 \cap B_2 \dots \cap B_n) = P(B_1)P(B_2)\dots P(B_n). \tag{4.22}$$

Je n'ai pas donné de définition très pratique de l'indépendance de n événements B_j, mais il devrait être clair que que les ensembles B_j sont indépendants, on doit avoir (4.22).

En fait la définition la plus pratique que je pourrais donner maintenant, serait que les ensembles B_j sont indépendants si et seulement si les variables aléatoires $\mathbb{1}_{B_j}$ sont indépendantes. C'est un peu tiré par les cheveux, mais ca marche (et je vous laisse vérifier que pour deux ensembles, on trouve en fait la même définition). Plus sérieusement, on peut aussi vérifier qu'avec la définition d'indépendance raisonnable (celle que j'ai donnée pour 3 ensembles), on peut démontrer que si les variables X_j sont indépendantes, alors pour tout choix de α comme ci-dessus, les ensembles $B_j = X_j^{-1}(\alpha_j) = \big\{\omega \in \Omega\,;\, X_j(\omega) = \alpha_j\big\}$ sont indépendants aussi. Comme c'est une vérification peu drôle, on ne la fait pas (ouf).

Parlons maintenant un peu de la loi jointe des X_j. On pose $X = (X_1, \dots, X_n)$, définie par $X(\omega) = (X_1(\omega), \dots, X_n(\omega))$, et qui est à valeurs dans E^n. C'est une variable aléatoire, et elle a elle aussi une loi (qui est ce que je vais appeler la loi jointe des X_j. Cette loi est une mesure sur E^n, ou si vous préférez sur le produit $X_1(\Omega) \times X_2(\Omega) \dots, X_n(\Omega)$, que a l'avantage d'être fini. Notons cette loi $Q = Q_X$ (la lettre P est déjà beaucoup utilisée). On a juste besoin de savoir qui est le germe de cette mesure de probabilité, à savoir

$$\begin{aligned} q(\alpha_1, \dots \alpha_n) = Q(\{(\alpha_1, \dots \alpha_n)\}) &= P\big(X(\omega) = (\alpha_1, \dots \alpha_n)\big) \\ &= P\big(X_1(\omega) = \alpha_1 \text{ et } X_2(\omega) = \alpha_2 \dots \text{ et } X_n(\omega) = \alpha_n\big) \end{aligned} \tag{4.23}$$

En effet, une fois qu'on a le germe q, on peut retrouver la mesure Q en additionnant les probabilités.

Proposition 4.7. *Soient $X_1, \dots X_n$ des variables aléatoires (de Ω dans un ensemble E), et notons $X = (X_1, \dots, X_n)$. Alors les variables aléatoires $X_1, \dots X_n$ sont indépendantes si et seulement si la loi de X est la mesure produit des lois des X_j.*

L'énoncé a un sens puisque la loi $Q = Q_X$ est définie sur le produit des $X_j(\Omega)$, et la loi de X_j est une mesure sur $X_j(\Omega)$. Pour démontrer la proposition, on remarque que la définition (4.21) d'indépendance des X_j dit exactement que

$$q(\alpha_1, \dots \alpha_n) = Q(\{(\alpha_1, \dots \alpha_n)\}) = \prod_{j=1}^{n} P\big(\big\{\omega \in \Omega\,;\, X_j(\omega) = \alpha_j\big\}\big) = \prod_{j=1}^{n} q_j(\alpha_j), \tag{4.24}$$

où l'on a noté $q_j(\alpha) = P\big(X_j = \alpha\big)$, ce qui signifie aussi que q_j est le germe de la loi de X_j. Mais (4.24) dit exactement que le germe de Q est donné par $q(\alpha_1, \dots \alpha_n) = q_1(\alpha_1) \dots q_n(\alpha_n)$, qui est exactement la formule pour le germe de la probabilité produit. Voir (3.13) pour le produit de 2 mesures, la remarque qui suit pour un produit de 3, et autrement procéder par récurrence sur le nombre de termes. Donc en cas d'indépendance, la formule (4.24) dit que la mesure de Q est la mesure produit, mais réciproquement aussi, si la loi de X est la mesure produit, alors on a (4.24), donc aussi (4.21) (pour tout choix des α_j, et les variables sont indépendantes. □

Voici maintenant une définition équivalente de l'indépendance, qui ressemble à la définition initiale.

Proposition 4.8. *Soient $X_1, \dots X_n$ des variables aléatoires (de Ω dans un ensemble E) indépendantes. Alors, pour tout choix d'ensembles $A_1, A_2, \dots A_n \subset E$,*

$$P\big(\{\omega \in \Omega\,;\, X_1(\omega) \in A_1, \dots, \text{et } X_n(\omega) \in A_n\}\big) = \prod_{j=1}^{n} P\big(\{\omega \in \Omega\,;\, X_j(\omega) \in A_j\}\big). \tag{4.25}$$

On connaît le cas particulier où $A_j = \{\alpha_j\}$ est un ensemble à un élément, et dans ce cas (4.23) est exactement (4.21). Donc (4.23) est plus générale, et le lemme dit bien qu'elle donne une définition équivalente de l'indépendance.

Curieusement, c'est plus pratique une fois qu'on a la proposition 4.7, parce que ceci va nous éviter des calculs. De toute manière, l'idée de la démonstration doit être de découper les ensembles A_j en morceaux, et de calculer les sommes.

On suppose donc que les X_j sont indépendantes, on se donne des ensembles A_j, et on veut démontrer (4.25). En notant encore Q la loi (jointe) de $X = (X_1, \dots, X_n)$, le membre de gauche est $Q(A_1 \times A_2 \dots \times A_n)$. Notons Q_j la loi de X_j, de sorte qu'on sait par la proposition 4.7 que Q est le produit des Q_j. On peut se souvenir directement (à moins qu'on ne l'ait fait que pour un produit de deux mesures) que

$$Q(A_1 \times A_2 \dots \times A_n) = Q_1(A_1) Q_2(A_2) \dots Q_n(A_n); \tag{4.26}$$

alors comme $Q_j(A_j) = P(X_j \in A_j)$ par définition de la loi Q_j de X_j, on en déduit bien (4.25). Si on ne se souvient plus de (4.26), on le re-démontre, en notant que

$$\begin{aligned} Q(A_1 \times A_2 \dots \times A_n) &= \sum_{(\alpha_1, \dots \alpha_n) \in A_1 \times A_2 \dots \times A_n} Q(\{(\alpha_1, \dots \alpha_n)\}) = \sum_{(\alpha_1, \dots \alpha_n) \in A_1 \times A_2 \dots \times A_n} q(\alpha_1, \dots \alpha_n) \\ &= \sum_{(\alpha_1, \dots \alpha_n) \in A_1 \times A_2 \dots \times A_n} q_1(\alpha_1) \dots q_n(\alpha_n) = \prod_{j=1}^{n} \Big\{ \sum_{\alpha_j \in A_j} q_j(\alpha_j) \Big\} = \prod_{j=1}^{n} Q_j(A_j), \end{aligned}$$

où si vous préférez vous pouvez partir de la fin et développer en une grande somme. En tout cas, on a bien prouvé (4.25) et la proposition. On aurait pu faire le calcul plus directement sans passer par les mesures produits, mais j'ai essayé et trouvé ceci un peu plus désagréable, et de toute caçon c'est bien de mesures produit qu'il s'agit ! □

Fin de Cours 5, 2020 (avec un petit retour au cours 6)

Avant de continuer avec les variables indépendantes, une petite remarque sur l'innocuité d'ajouter des variables supplémentaires, ou de remplacer un espace de base (Ω, P) par un autre plus gros obtenu en faisant un produit par autre chose.

Lemme 4.9. *Soient (Ω, P) un espace probabilisé, et X une variable aléatoire (de Ω dans un ensemble E). Donnons-nous aussi un autre espace probabilisé (Ω_2, P_2), puis notons $\widehat{\Omega} = \Omega \times \Omega_2$ et $\widehat{P} = P \otimes P_2$ la mesure produit sur $\widehat{\Omega}$. Notons encore $\widehat{X}$ la variable aléatoire sur $\widehat{\Omega}$ définie par $\widehat{X}(\omega, \omega_2) = X(\omega)$ (moralement, donc c'est juste a même variable, mais on l'a juste définie sur un espace plus grand). Alors la loi de $\widehat{X}$ est égale à la loi de X.*

C'est utile parce que parfois, pour diverses raisons, on a envie de considérer un espace Ω plus grand que nécessaire pour X. Par exemple, X concerne 3 lancers de dés, mais on a envie plus tard de faire une dizaine d'autres lancers. Le lemme dit juste qu'il ne se passe rien de dangereux quand on fait ça.

La démonstration se doit d'être facile. On note d'abord que $\widehat{X}(\widehat{\Omega}) = X(\Omega)$, donc les deux lois sont définies sur le même ensemble. Il ne reste plus qu'à voir que pour tout $x \in X(\Omega)$, on a bien $P(X = x) = \widehat{P}(\widehat{X} = x)$. Et en effet

$$\begin{aligned}\widehat{P}(\widehat{X} = x) &= \widehat{P}\big(\{(\omega, \omega_2) \in \widehat{\Omega}\,;\, \widehat{X}(\omega, \omega_2) = x\}\big) = \widehat{P}\big(\{(\omega, \omega_2) \in \widehat{\Omega}\,;\, X(\omega) = x\}\big) \\ &= \widehat{P}\big(\{\omega \in \Omega\,;\, X(\omega) = x\} \times \Omega_2\big) = P\big(\{\omega \in \Omega\,;\, X(\omega)x\}\big)P_2(\Omega_2) = P(X(\omega) = x)\end{aligned}$$

puisque $P_2(\Omega_2)$, et comme souhaité. □

On revient aux variables aléatoires indépendantes. Le moyen le plus simple de fabriquer des variables aléatoire indépendantes est de les faire vivre sur des espaces produits. Plus précisément, voici un énoncé qui permet de construire des v.a. indépendantes.

Proposition 4.10. *Soient $(\Omega_1, P_1) \dots, (\Omega_n, P_n)$ des ensembles finis munis chacun d'une mesure de probabilité. Dotons le produit $\Omega = \Omega_1 \times \dots \times \Omega_n$ de la mesure de probabilité produit $P = P_1 \otimes \dots \otimes P_n$. Donnons nous, pour chaque $j \in \{1, \dots n\}$, une variable aléatoire $X_j : \Omega_j \to E$. On note aussi X_j la v.a. définie sur Ω par $X_j(\omega_1, \dots, \omega_n) = X_j(\omega_j)$. Alors les v.a. $X_j : \Omega \to E$ sont indépendantes.*

La vérification est juste histoire de tester les définitions. On a un peu abusé des notations, parce qu'en principe on aurait plutôt dû appeler $\widehat{X}_j$ la variable définie sur Ω, donc poser $\widehat{X}_j(\omega_1, \dots, \omega_n) = X_j(\omega_j)$.

Le plus simple pour vérifier l'indépendance des $\widehat{X}_j$, c'est de revenir à la définition 4.6. On se donne donc des éléments $\alpha_1, \alpha_2, \dots \alpha_n \in E$, et on calcule le membre de gauche de (4.21), à savoir $P(A)$, où

$$\begin{aligned}A &= \big\{\omega = (\omega_1, \dots \omega_n) \in \Omega\,;\, \widehat{X}_1(\omega) = \alpha_1, \dots \text{et } \widehat{X}_n(\omega) = \alpha_n\big\} \\ &= \big\{\omega = (\omega_1, \dots \omega_n) \in \Omega\,;\, X_1(\omega_1) = \alpha_1, \dots \text{et } X_n(\omega_n) = \alpha_n\big\} \\ &= \big\{\omega_1 \in \Omega_1\,;\, X_1(\omega_1) = \alpha_1\big\} \times \dots \times \big\{\omega_n \in \Omega_n\,;\, X_n(\omega_n) = \alpha_n\big\}.\end{aligned}$$

Du coup, puisqu'on a pris la mesure produit sur Ω, il vient

$$P(A) = \prod_{j=1}^{n} P_j\big(\{\omega_j \in \Omega_j\,;\, X_j(\omega_j) = \alpha_j\}\big) = \prod_{j=1}^{n} P\big(\{\omega \in \Omega\,;\, X_j(\omega_j) = \alpha_j\}\big),$$

où on a utilisé pour la dernière égalité le fait que pour $A_j \subset \Omega_j$, $P_j(A_j) = P(\widehat{A}_j)$, où l'on note $\widehat{A}_j = \Omega_1 \times \dots \Omega_{j-1} \times A_j \times \dots \Omega_n$ la partie $\{w_j \in A_j\}$ de Ω qui correspond à $A_j \subset \Omega_j$. Donc on a démontré (4.21), et les v.a. X_j sont indépendantes. □

Encore une proposition générale sur les variables indépendantes. Il s'agit maintenant de montrer que des fonctions de variables indépendantes sont encore indépendantes. Pensez par exemple que vous avez tiré 3 dés de manière indépendante. Prenez les variables suivantes : X_1 est le tirage du premier dé, X_2 le cube du tirage du second dé, et X_3 la parité du troisième lancer. Ces trois v.a. sont indépendantes, comme on le voit dans l'énoncé suivant.

Proposition 4.11. *Soient $X_1, \dots X_n$ des variables indépendantes, allant de Ω dans des ensembles E_i. Soit, pour tout j, une fonction $f_j : E_j \to E$. Alors les variables $f_j \circ X_j : \Omega \to E$, sont indépendantes.*

Ici j'ai énoncé l'indépendance des X_j en supposant que $X_j : \Omega \to E_j$, alors que plus haut on avait pris directement $X_j : \Omega \to F$ pour un ensemble F qui ne dépend pas de j. J'ai fait ceci pour autoriser plus de liberté pour le choix des f_j, mais d'autres choix auraient été possibles, avec la même démonstration.

Une autre remarque avant la démonstration. Une fonction de X_j possible est la fonction constante. Ceci ne doit pas vous surprendre. Quand X est constante, ou juste un peu plus généralement quand il existe une constante C telle que $P(X = C) = 1$ (et donc $P(X \neq C) = 0$), X est automatiquement indépendante de toute autre variable aléatoire Y que vous auriez pu choisir. En effet, on doit vérifier que pour tout choix de α, β,

$$P(X = \alpha \text{ et } Y = \beta) = P(X = \alpha)P(Y = \beta).$$

Mais de deux choses l'une. Ou bien $\alpha \neq C$, la condition $X(\omega) = \alpha$ n'est vérifiée que sur un ensemble de mesure nulle, et alors les deux membres sont nuls, soit au contraire $\alpha = C$, la condition $X(\omega) = \alpha$ est vérifiée sur un ensemble de probabilité 1, et les deux membres sont égaux à $P(Y = \beta)$.

Maintenant on passe à la démonstration de le proposition. On doit vérifier que chaque fois qu'on remplace une des variables X_j par $f_j \circ X_j$, on préserve l'indépendance. Donc il suffit de vérifier ceci pour un seul remplacement (on peut remplacer les variables une par une), par exemple pour la première variable. Donc oublions l'indice 1 dans f_1 et vérifions qu'on peut remplacer X_1 par $f(X_1)$ (et garder l'indépendance).

On doit vérifier que pour tout choix de $\alpha_1, \alpha_2, \dots \alpha_n$,

$$P(f \circ X_1 = \alpha_1 \text{ et } X_2 = \alpha_2 \dots \text{ et } X_n = \alpha_n) = P(f \circ X_1 = \alpha_1)P(X_2 = \alpha_2) \dots P(X_n = \alpha_n).$$

Autrement dit, en posant $A_1 = f \circ X_1^{-1}(\alpha_1)$, $A_2 = X_2^{-1}(\alpha_2), \dots A_n = X_n^{-1}(\alpha_n)$, on veut que

$$P(A_1 \cap \dots \cap A_n) = P(A_1)P(A_2) \dots P(A_n). \tag{4.27}$$

On note $Z = f^{-1}(\alpha_1)$; alors $f \circ X_1^{-1}(\alpha_1)$ est l'union disjointe, pour $z \in Z$, des ensembles $X_1^{-1}(z)$. Donc du coup

$$A_1 \cap \dots \cap A_n = \bigcup_{z \in Z} \Big(X_1^{-1}(z) \cap A_2 \dots \cap A_n \Big),$$

et l'union est disjointe. Or par indépendance des X_j, on a bien que pour tout $z \in Z$,

$$\begin{aligned} P(X_1^{-1}(z) \cap \ldots \cap A_n) &= P(X_1 = z \text{ et } X_2 = \alpha_2 \ldots \text{ et } X_n = \alpha_n) \\ &= P(X_1 = z)P(X_2 = \alpha_2) \ldots P(X_n = \alpha_n) \\ &= P(X_1 = z)P(A_2) \ldots P(A_n). \end{aligned}$$

Il ne reste plus qu'à sommer sur $z \in Z$ et on obtient (4.27), dont on déduit bien l'indépendance de $f(X_1), X_2, \ldots X_n$. □

Maintenant on est prêt pour la vérification de (4.17) (sur la somme de variables indépendantes de Bernoulli). On l'a déjà démontré dans le cas particulier où $\Omega = \Omega_1 \times \ldots \Omega_n$ est un ensemble produit, et chaque X_j est une v.a. sur Ω_j dont la loi est Bernoulli de paramètre p. En fait, on a pris pour X_j la j-ième coordonnée dans le produit.

Passons au cas général. Soient donc $\widetilde{X}_1, \widetilde{X}_2, \ldots \widetilde{X}_n$ des variables indépendantes, dont la loi (de chacune) est une loi de Bernoulli de paramètre p. Posons $\widetilde{X} = (\widetilde{X}_1, \widetilde{X}_2, \ldots \widetilde{X}_n)$; on sait à cause de la proposition 4.7 que la loi de $\widetilde{X}$ est le produit des lois des X_j, donc le produit de n mesures de Bernoulli de paramètre p. Or ceci est par construction la loi de notre variable $X = (X_1, \ldots, X_n)$. Autrement dit, $\widetilde{X}$ et X ont la même loi. Et maintenant, la loi de la somme de nos variables aléatoires ne dépend que le la loi jointe des variables. Autrement dit, la loi de $\widetilde{X}_1 + \ldots + \widetilde{X}_n$ est égale à la loi de $X_1 + \ldots + X_n$. Cette dernière est une loi binomiale de paramètres n et p (c'est le cas particulier qu'on a démontré) ; on en déduit la même chose pour $\widetilde{X}_1 + \ldots + \widetilde{X}_n$.

Et pourquoi la moi de $X_1 + \ldots + X_n$ ne dépend que de la loi jointe des X_j ? C'est parce qu'on peut calculer (comme on l'a fait d'ailleurs dans la démonstration autour de (4.19)) la probabilité $P(X_1 + \ldots + X_n = k)$ à partir de toutes les probabilités $P(X_1 = \alpha_1 \text{ et } X_2 = \alpha_2 \ldots \text{ et } X_n = \alpha_n)$. □

Voici un corollaire dont je trouve la démonstration amusante et dans le même esprit.

Corollaire 4.12. *Soient S et T deux variables aléatoires indépendantes. On suppose que la loi de S est binomiale de paramètres m et p, et que la loi de T est binomiale de paramètres n et p (le même p). Alors la loi de $S + T$ est binomiale de paramètres $m + n$ et p.*

On va commencer par un cas particulier important, qui donne aussi une idée de ce qu'on veut faire. Comment construire un exemple ? On a une idée de comment faire une v.a. de loi binomiale, en ajoutant des v.a. indépendantes de loi Bernoulli. Alors on pose $\Omega = E_2^{m+n}$, muni de la probabilité produit $P = P_1 \otimes \ldots \otimes P_{m+n}$, où chaque P_j est en fait la même mesure de Bernoulli sur E_2 de paramètre p. C'est ce qu'on avait fait pour les sommes de Bernoulli ; ici on prend juste un peu plus de termes.

On pose $Y_j(\omega) = \omega_j$, la j-ème coordonnée, pour $1 \leq j \leq m + n$. Par le cas particulier de (3.16) fait plus haut dit que la loi de $\sum_{j=1}^{m+n} Y_j$ est bien une loi binomiale de paramètres $m+n$ et p. On pose aussi $S = \sum_{j=1}^{m} Y_j$, et je prétends que la loi de S est une loi binomiale de paramètres m et p. C'est presque pareil que précédemment ; la petite différence est qu'au lieu de regarder les X_j $(1 \leq j \leq m)$ comme définis sur E_2^m, ils sont définis sur le plus gros espace

E_2^{m+n}, muni de la probabilité produit. Mais le lemme 4.9 dit que les deux options donnent la même loi (ajouter un produit supplémentaire ne change rien). De même, $T = \sum_{j=m+1}^{m+n} Y_j$ a une loi binomiale paramètres n et p. Et finalement S et T sont indépendantes, à cause de la proposition 4.10. Donc dans le cas particulier des variables S et T, le corollaire est bien vrai.

Qu'en est-il maintenant des autres couples (S', T') de variables aléatoires indépendantes, où S' et T' ont des lois binomiales avec les mêmes paramètres que ci-dessous ? On peut maintenant utiliser le fait que la loi de $S' + T'$ ne dépend que de la loi (jointe) du couple (S, T), et pas de la réalisation particulière de S' ou T'. Je rappelle qu'on vient de voir ceci en fin de démonstration de (4.17), et que c'est un principe assez simple : on connaît toutes les probabilités des ensembles $\{S' = \alpha \text{ et } T' = \beta\}$, et on peut facilement en déduire les probabilités des ensembles $\{S' + T' = \gamma\}$. Par exemple en distinguant toutes les choix de (α, β) tels que $\alpha + \beta = \gamma$, puis en sommant les probabilités.

Or la loi de (S', T') est obtenue en partant de la loi de S' (connue), et celle de T' (connue), en faisant un produit (par indépendance, voir la proposition 4.7). Donc quand on remplace (S', T') par (S, T), on ne change pas la loi de $S' + T'$. On trouve donc la loi de $S + T$, calculée plus haut (ouf). □

Remarque complementaire possible (et je n'ai pas regardé où ça tombe le mieux) : loi marginale. Si $X = (X_1, X_2)$ est une v.a. (et on peut même prendre X_1 et X_2 à valeurs dans des $\mathbb{R}^m$), on peut calculer la loi de X_1 à partir de celle de X en "oubliant" la seconde variable. Donc P_1 (la loi marginale par rapport à X_1) est juste donnée par $P_1(\{k\}) = P(\{X_1 = x\})$. En termes de mesures, il s'agit de projeter une mesure P sur un produit $\Omega_1 \times \Omega_2$ sur la première composante Ω_1, en posant simplement

$$P_1(A) = P(A \times \Omega_2) \text{ pour } A \subset \Omega_1.$$

On vérifie sans peine que P_1 est une mesure de probabilité sur Ω_1. Quand P est la mesure produit $P_1' \otimes P_2'$, on trouve aisément que $P_1 = P_1'$ Et c'est d'ailleurs le saul cas où l'on peut calculer P à partir des deux mesures de proba marginales ; en général c'est plus compliqué que cela. Exactement comme le fait qu'on ne peut pas savoir la loi du couple (X_1, X_2) juste en connaissant les lois de X_1 et X_2.

4.6 Encore deux lois classiques : géométrique et hypergéométrique

On commence par **la loi géométrique de paramètre p**. C'est une première entorse à notre principe de ne considérer que des ensembles finis, parce qu'ici ce sera une mesure de probabilité sur $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$. Faisons un peu comme si on n'avait pas vu, et on justifiera mieux plus tard.

Définition 4.13. *La mesure géométrique de paramètre $p \in]0, 1]$ est la mesure de probabilité $P = \mathcal{G}(p)$ sur $\mathbb{N}^*$ définie par le fait que*

$$P(\{k\}) = p(1-p)^{k-1} \quad \text{pour tout entier } k \geq 1. \tag{4.28}$$

On prend bien $p > 0$, parce que sinon on attribuerait la probabilité 0 à tout le monde et ce ne serait pas une mesure de probabilité. Voir l'exemple fondamental de la pièce plus bas.

On est encore dans le cas où l'on peut définir une mesure de probabilité à partir d'un germe, ici l'application $k \to p(1-p)^{k-1}$. C'est juste que certaines sommes devront être remplacées par des séries. Par exemple, si on veut définir la mesure d'une partie A de $\mathbb{N}$, on devra maintenant écrire

$$P(A) = \sum_{k=1}^{\infty} \mathbb{1}_A(k) p(1-p)^{k-1} = \lim_{N\to+\infty} \sum_{k=1}^{N} \mathbb{1}_A(k) p(1-p)^{k-1} \tag{4.29}$$

où maintenant on somme une série (ou on prend la limite d'une suite de sommes partielles), et comme précédemment on n'ajoute le germe $p(1-p)^{k-1}$ que quand $k \in A$.

Regardons déjà ce qui se passe quand $A = \mathbb{N}^*$; je prétends que

$$\sum_{k=1}^{\infty} p(1-p)^{k-1} = 1. \tag{4.30}$$

Calculons la somme partielle $S_N = \sum_{k=1}^{N} p(1-p)^{k-1}$ posons $a = 1-p$ pour se ramener à des calculs connus. Alors $S_N = p\big[1 + (1-p) + \dots (1-p)^{N-1}\big] = p\big[1 + a + \dots a^{N-1}\big] = p\frac{1-a^N}{1-a} = 1 - a^N = 1 - (1-p)^N$. On a utilisé le fait que $a < 1$ puisque $p \neq 0$.

Fin du cours 6, 2020, mais sans dire comment on trouve P à artir du germe.

Donc $\lim_{N\to+\infty} S_N = 1$, ce qui justifie (4.30). Noter que (4.30) est la condition principale pour avoir un germe de probabilité, qui dit que $P(\mathbb{N}^*) = 1$ pour l'univers entier.

Dans le cas de (4.29), pour un ensemble $A \subset \mathbb{N}^*$, on a maintenant une série à termes positifs, qui sont dominés par les coefficients $p(1-p)^{k-1}$; les sommes partielles sont inférieures aux S_N, et donc la série converge, avec une somme inférieure à 1, par la théorie que vous connaissez des séries à termes positifs.

Par des manipulations classiques sur les séries, on aurait encore

$$P(A) = \sum_{j} P(A_j) \quad \text{quand } A \subset \mathbb{N}^* \text{ est l'union disjointe des } A_j. \tag{4.31}$$

Ceci vaut pour une union finie (comme précédemment), mais encore, au prix de manipulations supplémentaires un peu plus compliquées, pour des unions infinies.

Bref, on vient de (dire qu'on peut) vérifier que les formules (4.28) et (4.29) permettent de définir une mesure de probabilité sur $\mathbb{N}^*$, qu'on note $\mathcal{G}(p)$ et qu'on appelle mesure géométrique de paramètre p. On va voir l'exemple emblématique dans peu de temps, mais commencons par calculer l'espérance, cela peut toujours servir.

Proposition 4.14. *Si X est une variable aléatoire (sur un espace probabilisé (Ω, P)) qui a pour loi $\mathcal{G}(p)$, alors $E[X] = \frac{1}{p}$.*

Noter que puisque X peut prendre une infinité de valeurs entières avec une probabilité non nulle, Ω aussi est irrémédiablement infini. En plus, le calcul de $E[X]$ aussi doit faire intervenir des séries ! Je prétends que

$$E[X] = \sum_{k=1}^{\infty} kP(X=k) = \sum_{k=1}^{\infty} kp(1-p)^{k-1} = \frac{1}{p}, \tag{4.32}$$

où la première partie est une définition logique de l'espérance et où l'énoncé contient le fait que la série converge. On en reparlera dans un chapitre ultérieur.

Attention, il y a des variables aléatoires, par exemple sur $(\mathbb{N}^*, \mathcal{G}(p))$, dont l'espérance ne peut être calculée parce que la série $\sum_k X(k)p(1-p)^{k-1}$ diverge atrocement. Par exemple, si $p = 1/2$, $X(k) = 10^k$ donne une espérance infinie. Donc ici on a de la chance.

Hélas la démonstration utilise un argument sur les séries de fonctions (ici, des séries entières) qu'il ne serait pas raisonnable que je redémontre entièrement. Je le livre donc tel quel, pour relecture quand vous aurez les outils (si ce n'est pas déjà le cas). On part de la série entière $\sum_{k\geq 0} x^k$, qui est convergente pour tout $x \in]-1,1[$ (mais ici on s'intéresse à $x \geq 0$) : on sait que sa somme est $S(x) = \sum_{k\geq 0} x^k = \frac{1}{1-x}$ (et d'ailleurs ce serait encore vrai pour $x \in \mathbb{C}$ tel que $|x| < 1$). Mais on sait aussi, par la théorie des séries entières, que pour $-1 < x < 1$, la série peut être dérivée terme à terme. Evidemment, c'est un peu particulier aus séries entières dans l'intervalle ouvert de convergence (ou alors aux séries de fonctions dont la série des dérivées converge normalement, par exemple). On trouve que $S'(x) = \sum_{k\geq 0}(x^k)' = \sum_{k\geq 0} kx^{k-1}$. Donc en calculant S', $\frac{1}{(1-x)^2} = \sum_{k\geq 1} kx^{k-1}$ pour $x \in]-1,1[$ (le terme pour $k = 0$ est nul). Prendre $x = 1-p$ est autorisé, puisque $0 < p \leq 1$, et on trouve

$$\sum_{k=1}^{\infty} kp(1-p)^{k-1} = p\sum_{k=1}^{\infty} kx^{k-1} = p\frac{1}{(1-x)^2} = \frac{1}{p}$$

comme annoncé dans (4.32). □

Exemple emblématique. On lance une pièce (obéissant à la loi de Bernoulli de paramètre p), et on décide de continuer jusqu'à obtenir le premier résultat égal à 1. La variable aléatoire est le numéro du premier lancer où ceci se produit.

On va déjà regarder ce qui se passe quand on lance N fois (on prévoit N grand). L'avantage est qu'on peut se placer dans l'espace $\Omega_N = E_2^N$, muni de la probabilité P_N qui est le produit de N probabilités de Bernoulli de paramètre p.

On peut définir la variable aléatoire X_N qui est presque ce que l'on veut. Si on a fini par tirer un 1, on pose $X_N(\omega) = \inf\{k \geq 1\,;\, \omega_k = 1\}$, qui correspond bien à ce qu'on a dit. Mais das le cas qui reste, à savoir si $\omega = (0,0,\dots 0) = 0$, on n'a pas encore défini $X_N(\omega)$, et il n'y a pas de choix complètement logique. On peut poser $X_N(\omega) = N+1$ (mais on sait qu'on voulait autre chose), ou $X_N(\omega) = *$ (juste une lettre différente, pour indiquer qu'on n'est pas content, c'est un peu ennuyeux parce que $*$ n'est pas un nombre, mais ce n'est pas grave parce qu'on ne va pas calculer l'espérance de X_N). On espére surtout que comme $P(\{0\}) = (1-p)^N$ (et qu'on a pris $p > 0$), ce dernier cas est si peu probable qu'il n'aura pas d'influence.

La solution élégante, mais qu'on s'interdit dans ce cours, serait en fait de prendre un produit infini Ω_∞ de copies de E_2, y mettre une mesure de probabilité qui est le produit infini auquel on pense, et de voir tous nos Ω_N comme des projections de Ω_∞.

En attendant on peut commencer à calculer la loi de X_N. Appelons $q(k)$ la probabilité d'avoir $X_N = k$.

D'abord, $q(1) = p$: c'est la probabilité d'avoir gagné du premier coup. Du coup, la probabilité de ne pas avoir gagné en $k = 1$ est $1 - p$.

Maintenant $q(2)$ est la probabilité de ne pas avoir gagné au premier coup, à savoir $1 - p$, multipliée par la probabilité conditionnelle, sachant cela, de gagner au second coup. On trouve $q(2) = p(1 - p)$. Et du coup, la probabilité de ne pas encore avoir gagné est $(1 - p)^2$ (soit soustraire, soit calculer directement la probabilité de perdre les deux fois !). On a ici utilisé le fait que nos lancers de pièces sont (implicitement) supposés indépendants !

Je vous laisse vérifier en recommencant que $q(3) = p(1 - p)^2$ (on a perdu 2 fois, et sachant cela, on a gagné au troisième coup), et que la probabilité du reste est $(1 - p)^3$. Puis par récurrence que $q(k) = p(1 - p)^{k-1}$.

Evidemment si on est malchanceux on arrive jusque N, et la probabilité d'avoir raté les N lancers est $(1 - p)^N$ comme annoncé. A ce stade on peut se désoler de ne pas voir prévu assez de lancers (et d'être si malchanceux), mais on voit bien aussi que le calcul des $q(k)$, $k \leq N$, ne dépend pas de notre choix de N, et on peut toujours relancer le calcul jusque $2N$, calculer jusque là, et ainsi de suite. L'événement restant où l'on tire une infinité de 0 à la suite est de probabilité nulle (inférieure à $(1 - p)^N$ pour tout N), donc on peut bien en déduire que la loi de X (la variable correspondant à X_N, mais où l'on fait autant de lancers que ce dont on a besoin) est bien la loi $\mathcal{G}(p)$ de la définition. □

Remarque. On a en fait calculé en cours de route la probabilité de l'événement $\{X > k\}$ (on a perdu les k premières fois). On a trouvé $P(\{X > k\}) = (1 - p)^k$. Exercice : le vérifier soit par calcul direct (déjà fait, n'est-ce pas), soit à partir de la loi $\mathcal{G}(p)$.

Exercice (j'espère, sans application pratique). Un étudiant pas très brillant et sans mémoire à long teme passe l'examen de probastats tous les ans, avec une chance sur 10 de réussir, jusqu'à ce qu'il l'obtienne. Quelle est l'espérance du nombre d'années qu'il lui faudra pour obtenir le certificat ? Quelles sont les chances qu'il ne l'ait toujours pas après 20 tentatives ?

On passe à la **loi hypergéométrique de paramètres N, G, et n**.

Cette fois, donnons l'exemple pour commencer. Le nom est compliqué, mais ce sera assez simple. On se donne un amphi de N étudiants. Dans cet amphi, il y a G gauchers (et $N-G$ droitiers). On tire au hasard n étudiants parmi les N. Et on considère la variable aléatoire X qui est le nombre de gauchers tirés.

La loi de cette variable aléatoire X sera appelée loi hypergéométrique de paramètres N, G, et n. On va la calculer maintenant, mais en tout cas c'est une mesure de probabilité sur $\{0, 1, \dots n\}$.

Evidemment, on aurait pu trouver plein d'exemples équivalents (un lac avec deux espèces de poissons, on en pêche $n < N$, ou un éléctorat pour le second tour, et on fait un sondage). Mais tous donnant la même loi.

Attention (on en reparlera plus tard), ici on prend n étudiants différents, on ne remet pas les poissons, et on ne sonde pas deux fois la même personne. Les calculs seraient plus simples si on faisait ça.

Proposition 4.15. *La loi hypergéométrique de paramètres N, G, et n est donnée par la formule suivante : pour $0 \leq k \leq n$,*

$$P(X = k) = \frac{C_G^k C_{N-G}^{n-k}}{C_N^n}. \tag{4.33}$$

On notera $\mathcal{H}(N, G, n)$ cette loi.

Gardez les conventions sur le fait que $C_a^b = 0$ quand $b > a$. On en a besoin quand par exemple il y a strictement moins de k gauchers en tout ; alors il est logique que la probabilité d'en ramasser k soit nulle ! Souvent ceci ne se produit pas parce que n est choisi plus petit que N et G.

L'avantage de la formule ci-dessus est qu'elle est exacte. Son inconvénient est qu'elle est assez compliquée (mais pas tant).

Pour la démonstration, on modélise notre histoire d'étudiants dans l'amphi (ou tout autre problème équivalent). Le meilleur Ω possible est l'ensemble des parties de E_N qui ont exctement n éléments Noter que Ω a exactement C_N^n éléments, notre dénominateur. Il faudra donc voir que le nombre des éléments de Ω (des parties A de E_N à n éléments) pourlesquelles $X(A) = n$ (autrement dit, A contient k gauchers) est $C_G^k C_{N-G}^{n-k}$.

Et comment construit-on une partie A comme ceci ? On doit choisir k gauchers parmi les G gauchers, et $n-k$ droitiers parmi les $N-G$ droitiers. Les

deux choix sont indépendants (en fait on a juste besoin de savoir que pour tout choix des gauchers, le nombre de choix pour les droitiers est toujours le même). On a donc un nombre total de choix qui est le produit de C_G^k (pour les gauchers) et C_{N-G}^{n-k} (pour les droitiers). C'est ce qu'on a annoncé. □

Amusant pour moi : on sait que $\sum_{k=0}^{n} \frac{C_G^k C_{N-G}^{n-k}}{C_N^n} = 1$ puisque la loi est une mesure de probabilité. Probablement possible de le retrouver à partir de l'expression avec des factorielles, mais c'est moins rigolo.

Et a propos, pouvez vous deviner l'espérance ? La formule $E = \sum_{k=0}^{n} k \, \frac{C_G^k C_{N-G}^{n-k}}{C_N^n}$ ne semble pas aider trop, alors qu'on peut deviner la réponse.

Proposition 4.16. *Soit X une variable aléatoire dont la loi est hypergéométrique de paramètres N, G, et n. Alors $E[X] = \frac{nG}{N}$.*

Bon, ceci paraît logique, mais comment le démontrer ? Démonstration astucieuse (j'ai du copier). On se souvient de notre modèle, où Ω est l'ensemble des parties de E_N qui ont n éléments. Pour tout $j \in E_N$, on définit une variable aléatoire X_j, qui vaut 1 si j a été choisi dans la partie $\omega \in \Omega$, et 0 sinon. Et on pose $X = \sum_{j \in \mathcal{G}} X_j$, où $\mathcal{G}$ désigne l'ensemble des gauchers. Donc $X(\omega)$ est le nombre d'éléments de $\mathcal{G}$ qui se trouvent dans ω ; c'est exactement la variable qu'on avait choisie dont la loi est la loi hypergéométrique. Du coup,

$$E[X] = \sum_{j \in \mathcal{G}} E[X_j],$$

et heureusement le calcul de $E[X_j]$ (l'espérance d'avoir choisi j dans notre ω) est plus simple. Il faut juste regarder la proportion d'ensembles ω (à n éléments) qui contiennent j. Le nombre total est toujours C_N^n, et le nombre d'ensembles qui contiennent j est C_{N-1}^{n-1} (il reste $n-1$ éléments à choisir dans $E_N \setminus \{j\}$. Donc pour tout j,

$$E[X_j] = \frac{C_{N-1}^{n-1}}{C_N^n} = \frac{n}{N}$$

(calculez !) et en additionnant sur $j \in G$, $E[X] = \frac{nG}{N}$ comme promis. On s'en tire bien ! □

Exercice. Notons $\mathcal{H}(N, G, n)$ la loi hypergéométrique de paramètres N, G, et n. On va vérifier que $\mathcal{H}(N, G, n) = \mathcal{H}(N, n, G)$ (on a échangé le nombre de

gauchers et la taille de l'échantillon). D'abord, je vous laisse faire en écrivant $\frac{C_G^k C_{N-G}^{n-k}}{C_N^n}$ comme un produit, en faisant pareil avec $\frac{C_n^k C_{N-n}^{G-k}}{C_N^G}$ où on a échangé, puis en vérifiant que c'est pareil.

Vous vous doutez maintenant que ce n'est pas la vraie raison. En fait, dans cette histoire k est le nombre d'éléments de l'intersection de $\mathcal{G}$ avec ω, deux ensembles aléatoires (l'un avec G éléments, l'autre avec n éléments), l'histoire est donc plus symétrique qu'on avait dit. Mais c'est quoi cette histoire d'ensembles d'aléatoires, et on sent bien qu'on a besoin d'indépendance aussi.

On change un peu de modèle, pour décrire deux tirages aléatoires indépendants. Cette fois Ω est le produit de Ω_1, l'ensemble des parties de E_N à G éléments, et Ω_2, l'ensemble des parties de E_N à n éléments. Sur chacun on met la mesure uniforme, ce qui signifie que $P_1(\omega_1) = \frac{C_N^G}{2^N}$ pour tout $\omega_1 \in \Omega_1$ et $P_2(\omega_2) = \frac{C_N^n}{2^N}$ pour tout $\omega_2 \in \Omega_2$. Et on met la probabilité produit sur Ω (tirages indépendants). C'est notre droit de définir cet espace, et maintenant on va calculer la loi de la variable aléatoire X définie par $X(\omega) = \mathrm{Card}(\omega_1 \cap \omega_2)$ de deux manières différentes.

Pour chaque $\alpha \in \Omega_1$ (un choix des gauchers), on peut calculer la probabilité, sachant que $\omega_1 = \alpha$, d'avoir $X = k$. On voit que cette probabilité est donnée par la loi $\mathcal{H}(N, G, n)$ (et ne dépend pas de α, essentiellement parce que la loi en question ne dépendait pas du choix des gauchers dans E_N) donc vaut $\frac{C_G^k C_{N-G}^{n-k}}{C_N^n}$. Maintenant on utilise la formule de Bayes (version de (3.27), qui ressemble juste à une définition) :

$$P(X=k) = \sum_{\alpha \in \Omega_1} P(\omega_1 = \alpha \text{ et } X = k) = \sum_{\alpha \in \Omega_1} P(\omega_1 = \alpha) P(X = k \text{ sachant } \omega_1 = \alpha)$$
$$= \sum_{\alpha \in \Omega_1} P(\omega_1 = \alpha) \frac{C_G^k C_{N-G}^{n-k}}{C_N^n} = \frac{C_G^k C_{N-G}^{n-k}}{C_N^n} \sum_{\alpha \in \Omega_1} P(\omega_1 = \alpha) = \frac{C_G^k C_{N-G}^{n-k}}{C_N^n}$$

Mais on aurait pu faire le calcul en échangeant les variables : pour tout choix de $\beta \in \Omega_2$, la probabilité d'avoir $X = k$ sachant que $\omega_2 = \beta$ est donnée par la loi $\mathcal{H}(N, n, G)$, et vaut $\frac{C_n^k C_{N-n}^{G-k}}{C_N^n}$. Le même calcul que plus haut donne alors $P(X = k) = \frac{C_n^k C_{N-n}^{G-k}}{C_N^n}$, les deux nombres sont les mêmes pour tout k, et donc les deux lois coincident.

Maintenant comparons la loi $\mathcal{H}(N, G, n)$ avec une loi bininomiale. Reprenons notre exemple de poissons dans un lac, disons des gardons et des autres. On a décrit une pêche de n poissons parmi N, sans remise à l'eau, mais on aurait pu aussi pêcher n poissons, de manière indépendante, avec remise à l'eau. De sorte que chaque fois qu'on sort un poisson, on a la probabilité G/N que ce soit un gardon. La variable aléatoire $\widetilde{X}$, nombre de gardons sortis parmi n, est alors décrit par une loi binomiale de paramètres n et p, avec $p = G/N$. Est-ce que les lois de X et $\widetilde{X}$ sont proches ?

D'abord, noter que les valeurs prises par X et $\widetilde{X}$ sont toutes les deux dans $[0, n]$, ce qui est un bon signe. Les deux ont la même espérance nG/N, ce qui est bon signe aussi. Choisir une autre loi binomiale aurait été peu réaliste et aurait donné un autre intervalle ou une autre espérance.

Quand G et n sont du même ordre, par exemple, on ne s'attend pas à ce que les deux lois se ressemblent trop. Pour prendre un cas plus extrême, si $n > G$, $P(X = n) = 0$ (on ne prendra pas plus de gardons qu'il n'y en a dans le lac) alors que $P(\widetilde{X} = n) > 0$.

Mais ce qui nous intéresse, c'est le cas où n est petit par rapport à G (donc à N aussi). S'il y a plein de gardons dans le lac, en avoir prélevé au plus n ne change pas trop la probabilité d'en attraper un nouveau. Et dans ce cas $P(X = k)$ doit être proche de $P(\widetilde{X} = k)$. On va l'énoncer de manière formelle, plus pour donner une idée d'énoncé qui a un sens qu'autre chose.

Proposition 4.17. *On se donne $p \in [0, 1]$ (une proportion de gardons), et deux entiers $n \geq 0$ et $k \in [0, n]$. On se donne aussi deux suites $\{N_j\}$ et $\{G_j\}$ d'entiers, avec $0 \leq G_j \leq N_j$, et on suppose que*

$$\lim_{j\to+\infty} N_j = +\infty \qquad \text{et} \lim_{j\to+\infty} \frac{G_j}{N_j} = p. \tag{4.34}$$

Alors

$$\lim_{j\to+\infty} \frac{C_{G_j}^k C_{N_j-G_j}^{n-k}}{C_{N_j}^n} = C_n^k p^k (1-p)^{n-k}. \tag{4.35}$$

On aura reconnu à gauche de (4.35) le nombre $P(X = k)$ pour toute variable aléatoire X de loi $\mathcal{H}(N_j, G_j, n)$, et à droite le nombre $P(\widetilde{X} = k)$ pour toute variable aléatoire X de loi binomiale de paramètres n et p. Cette fois j'espère

qu'il est clair que le membre de droite est plus simple (et il ne dépend plus de G ni de N, juste de la proportion).

Le fond de l'histoire est que dans la situation ci-dessus, la probabilité à chaque coup de tirer un gardon supplémentaire est très proche de p, donc la loi du m-ième poisson est très proche d'une loi de Bernoulli de paramètre p, et ensuite on additionne. Mais comme ici on a calculé les probabilités avec des factorielles, on en profite pour faire un calcul explicite.

Fin du cours 7, 2020

Soient donc n, k, G_j, N_j comme ci-dessus, et notons $N = N_j$ et $G = G_j$ pour gagner de la place. On suppose d'abord que G et $N - G$ tendent tous les deux vers $+\infty$; on s'occupera des autres cas plus tard. On doit calculer le nombre Z du membre de droite, et on trouve

$$Z = \frac{C_{G_j}^k C_{N_j-G_j}^{n-k}}{C_{N_j}^n} = \frac{C_G^k C_{N-G}^{n-k}}{C_N^n} = \frac{G!}{k!(G-k)!}\frac{(N-G)!}{(n-k)!(N-G-n+k)!}\frac{n!(N-n)!}{N!}$$

au moins quand j est assez grand pour que $G_j \geq k$ et $N_j - G_j \geq n - k$ (on a dit que G_j et $N_j - G_j$ tendent tous les deux vers $+\infty$). On sort de ce produit le terme fixe $\frac{n!}{k!(n-k)!} = C_n^k$, qui va nous donner le coefficient du membre de droite de (4.35). Ensuite on sort

$$Z_1 = \frac{G!}{(G-k)!} = G(G-1)(G-2)\dots(G-k+1),$$

qu'on compare à $\widetilde{Z}_1 = G^k$. Noter que $Z_1/\widetilde{Z}_1$ est un produit de k termes qui tendent tous vers 1 (parce que G tend vers $+\infty$), donc $Z_1/\widetilde{Z}_1$ tend vers 1. Ensuite on sort

$$Z_2 = \frac{(N-G)!}{(N-G-n+k)!} = (N-G)(N-G-1)\dots(N-G-n+k+1),$$

qu'on compare à $\widetilde{Z}_2 = (N-G)^{n-k}$. A nouveau $Z_2/\widetilde{Z}_2$ est un produit de $n-k$ termes qui tendent tous vers 1 (parce que $N - G$ tend vers $+\infty$), donc $Z_2/\widetilde{Z}_2$ tend vers 1. Il nous reste

$$Z_3 = \frac{(N-n)!}{N!} = (N(N-1)\dots(N-n+1))^{-1}$$

qu'on compare à $\widetilde{Z}_3 = N^{-n}$, et à nouveau $Z_3/\widetilde{Z}_3$ tend vers 1. Donc en résumé

$$Z = C_n^k \widetilde{Z}_1 \widetilde{Z}_2 \widetilde{Z}_3 H,$$

avec H qui tend vers 1, et $\widetilde{Z}_1\widetilde{Z}_2\widetilde{Z}_3 = G^k(N-G)^{n-k}N^{-n} = (G/N)^k((N-G)/N)^{n-k}$, qui tend vers $p^k(1-p)^{n-k}$ parce qu'on a supposé que $G/N = G_j/N_j$ tend vers p et donc $(N-G)/N = 1-G/N$ tend vers $1-p$. Donc Z tend vers $C_n^k p^k(1-p)^{n-k}$, comme promis.

Il nous reste les cas bizarres où G_j et $N_j - G_j$ ne tendent pas tous les deux vers $+\infty$. Commençons par le cas où G_j ne tend pas vers $+\infty$. Ceci signifie que $p = 0$, parceque sinon (4.33) nous dirait que G_j tend vers $+\infty$, en prenant le produit des limites. Dans ce cas on veut prouver que le membre de gauche de (4.35) tend vers 1 si $k = 0$ et vers 0 sinon. Pour $k = 0$, $Z = \frac{C_G^0 C_{N-G}^n}{C_N^n} = \frac{C_{N-G}^n}{C_N^n}$. Ici on sait que $(N-G)/N = 1-G/N$ tend vers 1 (par (4.34) et puisque $p = 0$), et donc $\frac{C_{N-G}^n}{C_N^n} = \frac{(N-G)\ldots(N-G-n+1)}{N\ldots(N-n+1)}$ tend vers 1 par le même raisonnement que plus haut. Donc $Z = Z(0)$ tend vers 1 quand $k = 0$. Mais maintenant, la probabilité restante pour tous les autres k est $1-Z(0)$, qui tend vers 0, donc tous les $Z(k)$, $k > 1$, tendent vers 0.

Finalement il ne reste que le cas où $N_j - G_j$ ne tend pas vers $+\infty$. Comme plus haut, ceci implique que $p = 1$. Cette fois, on regarde le membre de droite de (4.35) et on voit qu'on doit montrer que Z tend vers 1 quand $k = n$ et tend vers 0 sinon. Quand $k = n$, on trouve $Z = Z(n) = C_G^n C_{N-G}^0 / C_N^n = C_G^n / C_N^n$. Maintenant G/N tend vers 1 puisque $p = 1$, et $Z(n)$ tend vers 1 comme ci-dessus. Finalement, pour les autres valeurs de k, on a encore $Z(k) \leq 1 - Z(0)$ (puisque $\sum_k Z(k) = 1$), qui tend vers 0. On a fait tous nos cas. □

Remarque C'est en fait un peu surprenant qu'on doive se donner tant de mal alors que l'intuition était si claire, que les deux probas à chaque moment de pêcher un gardon sont aussi proches qu'on veut. On pourrait d'ailleurs aussi formaliser tout ceci en suivant le processus de pêche, et en estimant la différence entre les probabilités conditionnelles (à tout moment) de pêcher un gardon ou autre chose, dans les deux processus de pêche avec remise, ou sans remise. Je crois que finalement les calculs seraient plus simple dans le cas où $0 < p < 1$. Le cas où il y a très peu de gardons (ou très peu du reste) serait peut-être un peu plus délicat, sans doute à traiter à part à nouveau.

Je ne peux pas m'empêcher de dire que l'on a donné un énoncé en termes de suite parce que c'est souvent plus facile à comprendre, mais en fait le vrai énoncé (et vous pouvez vérifier qu'on l'a démontré en fait) est un énoncé de limite : étant donnés n et k (comme plus haut), et $\varepsilon > 0$, il existe $N_0 \geq 1$ et $\tau > 0$ tels que si X est une variable aléatoire de loi $\mathcal{H}(N, G, n)$, et si $N \geq N_0$ et $|\frac{G}{N} - p| < \tau$, alors $|P(X = k) - C_n^k p^k (1-p)^{n-k}| \leq \varepsilon$. Si vous trouvez que tout compte fait c'est plus simple, en fait on est d'accord.

5 Espaces de probabilité dénombrables

5.1 Séries numériques et familles sommables

a. Séries à termes positifs. Jusqu'ici, on a presque pu s'en sortir avec des univers finis (sauf la loi géométrique), mais on sent bien qu'il peut être utile de considérer Ω un peu plus grand. On va faire le minimum, et considérer des ensembles (au plus) dénombrables, c.-à-d. en bijection avec (une partie de) $\mathbb{N}$. On va voir que la généralisation est possible sans trop d'ennuis, en remplaçant les arguments de sommation simple par des arguments sur les séries numériques.

On va utiliser divers résultats sur les suites et séries numériques, que je rappelle maintenant sans trop de démonstrations.

D'abord, une notion importante, la borne supérieure d'un sous-ensemble de $\mathbb{R}$. Je le mets sous forme de théorème parce que c'est important.

Théorème 5.1. *Soit $A \subset \mathbb{R}$ une partie non vide de $\mathbb{R}$. On suppose que A est majorée (c.-à-d. qu'il existe $M \in \mathbb{R}$ tel que $a \leq M$ pour tout $a \in A$. Alors A admet une borne supérieure, c.-à-d. un $S \in \mathbb{R}$ tel que d'une part $a \leq S$ pour tout $a \in A$, et d'autre part, pour tout $x < S$, il existe $a \in A$ tel que $a > x$.*

Autrement dit, S est le plus petit majorant de A. On le note

$$S = \sup_{a \in A} a = \sup \big\{ a \, ; \, a \in A \big\}.$$

Parfois, on note aussi $\sup_{a \in A} a = +\infty$ quand A n'est pas majoré.

C'est un théorème (avec des définitions dedans), ou alors c'est une partie de la définition de $\mathbb{R}$, suivant comment vous avez construit $\mathbb{R}$.

Maintenant, on va parler d'une manière, basée sur la borne supérieure, de calculer les sommes de séries à termes positifs.

Quelques notations pour le prochain énoncé. D'abord, si A est un ensemble, je noterai $\mathcal{P}_0(A)$ l'ensemble des parties finies de A. Donc bien sûr $\mathcal{P}_0(A) \subset \mathcal{P}(A)$ (avec égalité si A est fini).

Soit $\{u,\}_{n\in\mathbb{N}}$ une suite de nombres positifs. Pour chaque partie finie $J \in \mathcal{P}_0(\mathbb{N})$, on note $S_J = \sum_{n\in J} u_n$ la somme finie correspondante.

Proposition 5.2. *La série $\sum_{n=0}^{+\infty} u_n$ est convergente si et seulement si l'ensemble des sommes finies $S_J = \sum_{n\in J} u_n$, est majoré. Si c'est le cas, alors*

$$\sum_{n=0}^{\infty} u_n = \sup\big\{S_J\,;\, J \in \mathcal{P}_0(\mathbb{N})\big\}. \tag{5.1}$$

Donc le second membre est la borne supérieure de l'ensemble des sommes partielles S_J (où J est n'importe quel sous-ensemble fini de $\mathbb{N}$). Si on avait seulement pris les ensembles $\{0, 1, \dots N\}$, on aurait pris la borne supérieure des sommes partielles $\sigma_N = \sum_{n=0}^{N} u_n$. Dans ce cas précis, on dit juste que pour une série à termes positifs, il y a convergence si et seulement si les sommes partielles σ_N sont majorées (indépendamment de N). Et aussi que la limite de σ_N est aussi la borne supérieure des σ_N (parce que $\{\sigma_N\}$ est une suite croissante).

Dans la proposition on dit que pour le même prix, on peut aussi prendre des sommes partielles qui correspondent à des ensembles finis quelconques, sans avoir à prendre les termes dans l'ordre (voir plus bas aussi). C'est assez facile à vérifier. D'abord, si les S_J sont majorées par un $M \in \mathbb{R}$, alors les σ_N aussi, puisque chaque σ_N est un exemple de S_J. Mais réciproquement, si les σ_N sont majorées par un S', alors les S_J aussi : prendre un S_J, ajouter tous les termes oubliés (combler les trous), et on trouve un σ_N qui est plus grand que S_J. Donc $S_J \leq \sigma_N \leq S'$. On en déduit déjà la première partie. Mais aussi que tout majorant des S_J est aussi un majorant des σ_N, et aussi réciproquement. Donc la famille des S_J a les mêmes majorants que celle des σ_N ; on en déduit qu'elles ont aussi la même borne supérieure. (Exercice : traduisez en argument avec les quantificateurs comme dans la définition). □

Corollaire 5.3. *On suppose que la série à termes positifs $\sum_{n=0}^{+\infty} u_n$ est convergente, et on note S sa somme. On se donne une bijection $\psi : \mathbb{N} \to \mathbb{N}$. Alors la série à termes positifs $\sum_k u_{\psi(k)}$ est convergente, et sa somme est S.*

Bref, on peut sommer dans n'importe quel autre ordre. Demander une bijection sert à ne pas sommer deux fois le même terme, et à ne pas en oublier non plus.

Vous avez sans doute vu ceci, peut-être avec une autre démonstration. On note que les sommes finies correspondant aux deux séries sont en fait les mêmes, donc en appliquant la proposition, les sommes finies sont majorées si et seulement si chacune des deux séries a des sommes partielles majorées, et aussi si c'est le cas, les sommes sont les mêmes. On a juste dit que quand on regarde les sommes finies, l'ordre des termes n'intervient plus !

b. Familles sommables à termes positifs.

Pour passer à la sommation par paquets, on va généraliser un peu et remplacer $\mathbb{N}$ par un ensemble (quelconque, mais on pense infini !) I d'indices.

Ce qui va remplacer la série, c'est une famille $\{u_i\}$, $i \in I$ de nombres pour l'instant supposés positifs.

On dit que les nombres positifs u_i, $i \in I$, forment une <u>famille sommable</u> quand la famille des $S_J = \sum_{i \in J} u_i$, où J est dans l'ensemble $\mathcal{P}_0(I)$ des parties finies de I, est majorée. Donc s'il existe $M \in \mathbb{R}$ tel que $S_J \leq M$ pour tout $J \in \mathcal{P}_0(I)$. Alors la somme de la famille, naturellement notée $\sum_{i \in I} u_i$, est la borne supérieure de l'ensemble des sommes S_J, $J \in \mathcal{P}_0(I)$. En bref,

$$\sum_{i \in I} u_i = \sup \Big\{ \sum_{i \in J} u_i \, ; \, J \in \mathcal{P}_0(I) \Big\}. \tag{5.2}$$

On pourra encore utiliser (5.2) quand les sommes ne sont pas majorées (et donc $\sum_{i \in I} u_i = +\infty$), mais on ne dira pas que la famille est sommable ! Vous notez bien que c'est pareil que plus haut, sauf qu'on a remplacé $\mathbb{N}$ par I.

De manière amusante, on n'a pas à supposer que I est dénombrable pour cette définition (on pourrait prendre $I = \mathbb{R}$), mais si la famille $\{u_i\}$, $i \in I$ est sommable, on peut démontrer que l'ensemble des $i \in I$ tels que $u_i \neq 0$ est au plus dénombrable. Passons à la sommation par paquets.

Proposition 5.4. *On se donne une famille sommable à termes positifs u_i, $i \in I$ (vous pouvez penser que $I = \mathbb{N}$ et que c'est une série), et une partition de I en ensembles A_ℓ, $\ell \in L$. Donc $I = \cup_{\ell \in L} A_\ell$ (union disjointe). Alors chacune des familles $u_i, i \in A_\ell$ est sommable, la famille des sommes $S_\ell = \sum_{i \in A_\ell} u_i$, $\ell \in L$, est sommable aussi, et en plus*

$$\sum_{i \in I} u_i = \sum_{\ell \in L} S_\ell = \sum_{\ell \in L} \Big\{ \sum_{i \in A_\ell} u_i \Big\}. \tag{5.3}$$

On a pas mal utilisé ceci pour des sommes finies, et je dis que ceci reste vrai pour des sommes infinies (mais majorées), à termes positifs. Je répète bien "à termes positifs" tout le temps, parce que sinon, sans hypothèse correcte pour compenser (typiquement, la famille des $\sum_n |u_n|$ est majorée), tous les résultats ci-dessus sont faux.

C'est intéressant même dans le cas où L est fini, mais c'est presque aussi facile à démontrer dans le cas que j'ai énoncé. Pour que l'énoncé contienne aussi le cas fini, autorisons certains des sous-ensembles A_ℓ à être vides.

Et, pour le cas où on en aurait besoin de manière cachée, prendre certains A_ℓ vides ne pose pas de problème : on utilise la convention qu'une somme vide (avec 0 termes) est nulle, et tout se passe bien.

La démonstration n'est pas trop dure, mais je ne donne qu'une idée : la somme $\sum_{i \in I} u_i$ est une borne supérieure de sommes finies ; chaque somme finie peut être décomposée comme ci-dessus ; on écrit consciencieusement les définitions avec les bornes supérieures et on s'aperçoit qu'on est passé à la limite (en prenant les sup). □

c. Familles sommables à termes réels.

Celles-ci seront l'analogue des séries absolument convergentes. Mais on voit les choses de manière un peu différente : on souhaite que l'ordre des termes compte peu, et on aime bien sommer par paquets.

On dit que la famille $\{u_i\}$, $i \in I$ à termes réels ($u_i \in \mathbb{R}$ pour tout i) est sommable quand la famille des valeurs absolues $|u_i|$ est sommable. Autrement dit, quand $\sum_{i \in I} |u_i| < +\infty$ ou quand il existe $M \in \mathbb{R}$ tel que $\sum_{i \in J} |u_i| \leq M$ pour tout $J \subset I$ fini.

Pour calculer la somme, on va se ramener au cas des termes positifs. Pour

tout $x \in \mathbb{R}$, posons

$$x_+ = \max(0, x) \quad \text{et} \quad x_- = \max(0, -x); \tag{5.4}$$

ce sont la "partie positive" et la "partie négative" de x (choisis pour être positifs tous les deux). On vérifie aisément que

$$x = x_+ - x_- \quad \text{et } |x| = x_+ + x_-. \tag{5.5}$$

Donc par exemple $u_i = (u_i)_+ - (u_i)_-$ et $|u_i| = (u_i)_+ + (u_i)_-$. On appelle somme de la famille sommable $\{u_i\}$, $i \in I$ à termes réels le nombre

$$\sum_{i \in J} u_i = \sum_{i \in J} (u_i)_+ - \sum_{i \in J} (u_i)_- \tag{5.6}$$

qui est bien définie parce que les deux sommes à termes positifs, qui sont dominées par $\sum_{i \in I} |u_i|$, sont finies. Une autre manière de dire est qu'on somme tous les termes positifs, puis on somme tous les termes négatifs (et ceci donne deux nombres finis parce qu'on a supposé la famille sommable), et à la fin on fait la soustraction des deux sommes.

C'est logique si on veut que la somme de se comporte bien vis-à-vis des additions et soustraction. En fait, avec un peu de travail, on vérifie que si les deux familles $\{u_i\}$ et $\{v_i\}$, $i \in I$, sont sommable, alors $\{u_i + v_i\}$, $i \in I$, est aussi sommable, et

$$\sum_{i \in J} (u_i + v_i) = \sum_{i \in J} u_i + \sum_{i \in J} v_i. \tag{5.7}$$

Le travail consiste juste à couper I en six, suivant les signes de u_i, v_i, et $u_i + v_i$, et à ajouter les morceaux en suivant bien les signes.

On vérifie aussi facilement que pour tout réel, $\sum_{i \in J} (\lambda u_i) = \lambda \sum_{i \in J} u_i$.

Et aussi que si $I = \mathbb{N}$, la famille $\{u_i\}$, $i \in I$, est sommable si et seulement si la série $\sum_i |u_i|$ est convergente (on dit que $\sum_i u_i$ est absolument convergente), et alors la somme est la même que la somme de la série. On déduit assez facilement cette équivalence (que je ne démontre pas pour gagner du temps), la version "série absolument convergente" suivante du corollaire 5.3.

Corollaire 5.5. *On suppose que la série à termes réels $\sum_{n=0}^{+\infty} u_n$ est absolument convergente, et on note S sa somme. On se donne une bijection $\psi : \mathbb{N} \to \mathbb{N}$. Alors la série $\sum_k u_{\psi(k)}$ est absolument convergente, et sa somme est S.*

On n'en aura pas besoin, je crois, mais la démonstration est simple et se fait comme pour le corollaire 5.3. Si on retire les mots "absolument", le résultat devient notoirement faux ; vous savez sans doute que la série $\sum_{n=1}^{\infty} \frac{(-1)^n}{n}$ est convergente (pas absolument), et qu'on peut changer l'ordre des termes pour obtenir une série dont la somme est n'importe quel nombre réel, ou alors qui ne converge pas.

Chez les familles sommables, on a fait l'hypothèse forte tout de suite, et on obtient l'invariance par renumérotation juste dans la définition, et la sommation par paquets comme suit :

Proposition 5.6. *On se donne une famille sommable à termes réels u_i, $i \in I$, et une partition de I en ensembles A_ℓ, $\ell \in L$. Alors chacune des familles $u_i, i \in A_\ell$ est sommable, la famille des sommes $S_\ell = \sum_{i\in A_\ell} u_i$, $\ell \in L$ est sommable aussi, et en plus*

$$\sum_{i\in I} u_i = \sum_{\ell\in L} S_\ell = \sum_{\ell\in L} \Big\{ \sum_{i\in A_\ell} u_i \Big\}. \tag{5.8}$$

C'est le même énoncé que pour la proposition 5.4. La démonstration est simple : on écrit $u_i = (u_i)_+ - (u_i)_-$ comme ci-dessus, on applique la proposition et ceci se démontre directement 5.4 à chacune des deux familles sommables à termes positifs $\{(u_i)_+\}$ et $\{(u_i)_-\}$ (avec la même partition que dans l'énoncé), et à la fin on regroupe les termes en utilisant (5.7). Je vous laisse faire.

On utilisera cette proposition quand on calcule les espérances de variables aléatoires à valeurs réelles.

Noter que ce qu'on a dit pour $u_i \in \mathbb{R}$ marcherait encore pour $u_i \in \mathbb{C}$; on découperait d'abord u_i en partie réelle et partie imaginaire, puis on travaillerait sur chacun des morceaux séparément. Pareil pour des familles à valeurs dans $\mathbb{R}^n$.

Fin du cours 8, 2020

5.2 Mesures de probabilité sur un ensemble Ω dénombrable

On a déjà fait le cas où Ω est fini, et on veut juste passer au cas dénombrable ; mais ce qu'on va faire est encore valable quand Ω est fini.

On pourrait se contenter du cas où $\Omega = \mathbb{N}$, en utilisant une bijection de $\mathbb{N}$ dans Ω pour numéroter ses éléments. C'est juste que ce que l'on va raconter ne dépend pas de la manière dont on numérote (d'où les remarques ci-dessus sur l'ordre des termes dans les séries à termes positifs). Prendre $\Omega = \mathbb{N}$, c'est d'ailleurs ce qu'on fera dans les exemples.

On garde les mêmes définitions pour les événements élémentaires ($\omega \in \Omega$) et les événements ($A \subset \Omega$). On définit encore un germe de probabilité comme une fonction $p : \Omega \to [0, 1]$ telle que la famille $\{p(\omega)\}$, $\omega \in \Omega$ soit sommable, et

$$\sum_{\omega \in \Omega} p(\omega) = 1. \tag{5.9}$$

Donc maintenant $\sum_{\omega\in\Omega} p(\omega) = \sup_{J\in\mathcal{P}_0(\Omega)} \sum_{\omega\in J}$ en notant encore $\mathcal{P}_0(\Omega)$ l'ensemble des parties finies de Ω.

Quand on a un germe de probabilité p sur Ω, on peut définir une mesure de probabilité P sur Ω, en posant

$$P(A) = \sum_{\omega \in A} p(\omega) \tag{5.10}$$

pour tout $A \subset \Omega$. La famille est sommable (pensez la série converge) parce que toutes les parties finies J de A sont des parties finies de Ω, donc $\sum_{\omega\in J} p(\omega) \leq 1$ par (5.8). Ceci ressemble fort à (3.2).

Maintenant la mesure de probabilité P a un certain nombre de propriétés faciles, qui sont les mêmes que plus haut (voir (3.3)-(3.5)), que je rappelle parce que c'est facile à recopier :

$$0 \leq P(A) \leq P(\Omega) = 1 \quad \text{pour tout } A \in \mathcal{P}(\Omega) \tag{5.11}$$

comme (3.3) (et je peux ajouter que $P(\emptyset = 0)$) ; on vient de le voir, et par la même démonstration

$$P(A) \leq P(B) \quad \text{pour tout choix de } A, B \in \mathcal{P}(\Omega) \text{ tels que } A \subset B \tag{5.12}$$

et notre analogue de (3.5)

$$P(A) = \sum_j P(A_j) \quad \text{quand } A \text{ est l'union } \underline{\text{disjointe}} \text{ des } A_j \in \mathcal{P}(\Omega)\,; \tag{5.13}$$

cette fois, surtout si on veut se permettre une union disjointe infinie, on doit juste utiliser la proposition 5.4 au lieu du résultat "trival" sur les ré-écritures de sommes. Je rappelle que le fait que

$$P(A^c) = 1 - P(A) \tag{5.14}$$

s'en déduit aussitôt, puisque Ω est l'union disjointe de A et A^c ; de même, le fait que

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{5.15}$$

se démontre à partir de deux partitions simples.

Je recopie aussi la définition d'une mesure de probabilité sur Ω (dénombrable ou fini)

Définition 5.7. *Une mesure de probabilité sur Ω, est juste une fonction $P : \mathcal{P}(\Omega) \to [0,1]$ qui vérifie les propriétés (5.11), (5.12), et (5.13).*

Pour écrire les sommes, on utilise à nouveau les rappels ci-dessus, et le fait que tous nos nombres à sommer sont positifs (et qu'en fait toutes les sommes écrites sont ≤ 1).

Enfin, on retrouve p si l'on veut à partir de P, en posant $p(\omega) = P(\{\omega\})$; le fait que ce soit un germe vient de (5.11) et (5.13). Et d'ailleurs je vous laisse vérifier que (5.12) se déduit aussi de (5.11) et (5.13).

Bon, après cette définition on poursuit l'étude faite plus haut, mais dans le cadre des Ω (finis ou) dénombrables. Dans les démonstrations, on est souvent amené à sommer des séries au lieu d'écrire des sommes finies, mais je vous promets que tout se passe bien. Donc je vous passe les détails.

Evidemment, pas de mesure uniforme sur $\mathbb{N}$ ou $\mathbb{N}^*$, mais par contre on a déjà vu une mesure sur $\mathbb{N}^*$, celle qui donne la loi géométrique $\mathcal{G}(p)$, donc définie par $p(k) = p(1-p)^{k-1}$; voir la définition 4.13 et vérifier qu'on n'a pas menti dans ce paragraphe.

5.3 Variables aléatoires (sur (Ω, P) avec Ω dénombrable).

Définition pareille qu'avant : une v.a., c'est une fonction X définie sur Ω, et le plus souvent à valeurs dans $\mathbb{R}$ (mais on a vu des exemples à valeurs dans $\mathbb{R}^n$).

Maintenant $X(\Omega)$ est au plus dénombrable, donc on peut y définir une mesure de probabilité à partir d'un germe. On pose donc, pour tout $x \in X(\omega)$,

$$p_X(x) = P(\{\omega \in \Omega\,;\, X(\omega) = x\}) = P(X = x) \tag{5.16}$$

(où dans la seconde expression on abuse un peu des notations, mais pas plus que d'habitude). Si par exemple $X : \Omega \to \mathbb{R}$, on se permet aussi de définir $p_X(x)$ par (5.16) quand $x \in \mathbb{R} \setminus X(\Omega)$; on trouve $p_X(x) = 0$, ce point ne comptera pas dans la loi, justement parce que $p_X(x) = 0$, mais on peut avoir envie de définir $p_X(x)$ pour tout x, par exemple quand on a du mal à calculer l'image $X(\Omega)$. De toute façon, tout ce qu'on somme et qui vaut 0 donne 0.

On vérifie maintenant que p_X est un germe de probabilité. C'est pareil qu'avant :

$$\sum_{x \in X(\Omega)} p_X(x) = \sum_{x \in X(\Omega)} P(X^{-1}(x)) = 1$$

puisque les $P(X^{-1}(x)$ sont une partition de Ω ; bien sûr maintenant on a probablement une infinité de termes, donc on utilise la proposition 5.4.

Comme plus haut, on définit une mesure sur $X(\Omega)$ à l'aide de ce germe. On appelle cette mesure la loi de X ; c'est donc une mesure P_X sur $X(\Omega)$ (au plus dénombrable), telle que

$$P_X(\{x\}) = P(X = \omega) \tag{5.17}$$

et plus généralement, en prenant une union disjointe au plus dénombrable,

$$P_X(A) = P(X \in A) \quad \text{pour } A \subset X(\Omega) \tag{5.18}$$

Exemple, donc, la mesure géométrique (la loi $\mathcal{G}(p)$) rappelée plus haut. Et bientôt la loi de Poisson.

5.4 Espérance de variables aléatoires positives (Ω dénombrable).

Ah, maintenant, une petite différence, avec l'espérance. En fait il va s'agir de sommer des séries, et comme dans les cas que vous connaissez, le résultat ne sera pas toujours défini. On commence avec les v.a. à valeurs positives.

Définition 5.8. *On se donne $X : \Omega \to [0, +\infty)$. On appelle espérance de X le nombre*

(5.19)
$$E[X] = \sum_{\omega\in\Omega} X(\omega)P(\{w\}) = \sum_{x\in X(\Omega)} xP(X = x) = \sum_{x\in X(\Omega)} xp_X(x) \in [0, +\infty].$$

La dernière expression est juste une ré-écriture de la seconde. Mais attention, on ne sait pas si les deux premières sommes convergent ; on dit juste que si l'une est finie, alors l'autre aussi, et alors les sommes sont égales. Et l'on dira que X est d'espérance finie, ou intégrable, quand $E[X] < +\infty$.

Il y a donc une vérification à faire. Supposons d'abord que $\sum_{\omega\in\Omega} X(\omega)P(\{w\}) < +\infty$, et notons S la somme. Soit A n'importe quelle partie finie de $X(\Omega)$. La somme $\sum_{x\in A} xP(X = \omega)$ est bien définie (chacun des termes est bien défini, puisqu'on a défini la loi), et en plus

$$\sum_{x\in A} xP(X = \omega) = \sum_{x\in A} \Big\{ \sum_{\omega\in X^{-1}(x)} P(\{\omega\}) \Big\} = \sum_{x\in A} \Big\{ \sum_{\omega\in X^{-1}(x)} X(\omega)P(\{\omega\}) \Big\}$$

où la somme intérieure a peut-être une infinité de termes, mais alors la série converge. On peut regrouper les termes (tous positifs), et reste $\sum_{x\in\Omega; X(\omega)\in A} X(\omega)P(\{\omega\})$, qui est inférieur à S puisque S correspond à toute la somme sans demander que $X(\omega) \in A$. Donc la seconde somme de (5.19) converge et donne au plus S.

Dans l'autre sens c'est un peu pareil. On suppose que la somme de droite converge, et on note S' sa somme. On se donne une partie finie de B de Ω, en on veut contrôler $S(B) = \sum_{\omega\in\Omega} X(\omega)P(\{w\})$. On regroupe suivant les images $x = X(\omega), \omega \in B$ (il y en a un nombre fini), et $S(B) = \sum_{x\in X(B)} \sum_{\omega\in B\cap X^{-1}(x)} xP(\{\omega\})$. On ajoute des termes à chaque image réciproque, et il vient

$$S(B) \leq \sum_{x\in X(B)} \sum_{\omega\in X^{-1}(x)} xP(\{\omega\}) = \sum_{x\in X(B)} xP(X = x),$$

qui a moins de termes que le second terme de (5.19). Donc $S(B) \leq S'$ pour tout $B \subset \Omega$ fini. On prend la borne sup et on trouve que $\sum_{\omega\in\Omega} X(\omega)P(\{w\})$ est fini et inférieur à S'. Je vous laisse vérifier qu'on a tout démontré dans (5.19). □

Exemple. Si X est une v.a. dont la loi est la loi géométrique de $\mathcal{G}(p)$, $0 < p \leq 1$, on a vérifié ci-dessus (en disant qu'on s'occuperait du coté dénombrable de Ω) que $E[X] = 1/p$. Je vous laisse vérifier qu'on n'a pas menti.

Contrexemple. Définissons une probabilité P sur $\mathbb{N}$ par $p(n) = 2^{-n-1}$ pour $n \geq 0$. Noter que $\sum p(n) = 1$, comme souhaité pour un germe. Prenons $X(n) = 4^n$. Alors $E[X] = \sum_{n\geq 0} X(n)p(n) = \sum_{n\geq 0} 2^{n-1} = +\infty$. C'est le genre de choses qui arrivent quand les variables aléatoires sont trop grandes à l'infini. C'est pourquoi il faudra prendre des précautions concernant les v.a. à valeurs dans $\mathbb{R}$.

Quelques dernières remarques. Quand X ne prend qu'un nombre fini de valeurs (quand $X(\Omega)$ est fini), la somme de droite est finie, même si pour chaque terme $P(X = x)$, on a peut-être dû sommer sur une infinité de petits ω. La vérité est que ceci ne change pas grand-chose, à part que la loi de X est sur un univers fini.

Si $X : \Omega \to (-\infty, 0]$ est négative, on peut définir $E[X] = -E[-X]$ sans problème. C'est quand les variables changent de signe qu'il faudra se méfier un peu.

5.5 Espérance de variables aléatoires réelles (Ω dénombrable).

Passons aux v.a. à valeurs réelles. Voici la bonne définition qui fait marcher les choses.

Définition 5.9. *On se donne donc Ω (dénombrable ; fini est trop simple), P une mesure de probabilité sur Ω, et $X : \Omega \to \mathbb{R}$ une v.a. On dira que X est intégrable, ou d'espérance finie, quand l'une des conditions équivalentes suivantes est satisfaites :*

$$E[|X|] < +\infty \tag{5.20}$$

$$\sum_{\omega\in\Omega} |X(\omega)|P(\{w\}) < +\infty \tag{5.21}$$

$$\sum_{x\in X(\Omega)} |x|P(X = x) < +\infty. \tag{5.22}$$

L'équivalence des trois est facile ; appliquer (5.19) à la v.a. $\omega \mapsto |X(\omega)|$. Si on applique bêtement, à la place de (5.22) on trouve $\sum_{x\in|X|(\Omega)} xP(|X| = x)$, ce qui

est pareil que ce qu'on obtient à partir de l'expression de (5.22) en regroupant les termes pour x et $-x$.

On va pouvoir maintenant définir l'espérance de X quand X est comme ci-dessus. Si on ne suppose pas X intégrable, on ne peut pas utiliser les familles sommables pour calculer la somme ci-dessous, et essayer avec des séries non absolument convergentes serait bien dangereux à cause de l'absence d'invariance par renumérotation.

Définition 5.10. *Soient Ω (dénombrable ; fini est trop simple), P une mesure de probabilité sur Ω, et $X : \Omega \to \mathbb{R}$ une v.a.. On suppose que $|X|$ est intégrable (comme ci-dessus). Alors on définit l'espérance $E[X]$ par*

$$E[X] = E[X_+] - E[X_-] = \sum_{\omega\in\Omega} X(\omega)P(\{w\}) = \sum_{x\in X(\Omega)} xP(X = x). \tag{5.23}$$

Ici la différence $E[X_+] - E[X_-]$ a un sens puisque puisque les deux termes sont finis, et la première identité est la définition de la somme de la famille sommable $X(\omega)P(\{w\})$, à partir de la partie positive et la partie négative. La seconde égalité s'obtient comme précédemment, en écrivant que pour chaque $x \in X(\Omega)$, $P(X = x)$ est la somme des $P(\{\omega\}$, $\omega \in X^{-1}(\omega)$.

Et maintenant quelques propriétés des l'espérance.

D'abord, un exemple trivial mais qui sert tout le temps : si $C \in \mathbb{R}$, la v.a. constante égale à C est intégrable, et son espérance est C. Quand $C \geq 0$, c'est une conséquence immédiate de (5.19), et quand $C < 0$ on utilise le fait que $E[-C] = -E[C]$.

Ensuite, la propriété d'espace vectoriel et la linéarité : Si X et Y sont deux v.a. intégrables sur (Ω, P), alors $X + Y$ est intégrable (exercice ; utiliser le fait que $|X + Y| \leq |X| + |Y|$), et

$$E[X + Y] = E[X] + E[Y]; \tag{5.24}$$

c'est une conséquence immédiate du même résultat pour les familles sommables ; voir (5.7). De même, pour $\lambda \in \mathbb{R}$, λX est intégrable et $E[\lambda X] = \lambda E[X]$.

Puis les majorations : Si X et Y sont deux v.a. intégrables sur (Ω, P), et si $X(\omega) \leq Y(\omega)$ pour tout $\omega \in \Omega$, alors $E[X] \leq E[Y]$. Vrai sans restriction d'intégrabilité pour X et Y positives ; ensuite, dans le cas intégrable, appliquer (5.24) à $Y - X$ et noter que si $Y - X \geq 0$, alors $E[Y - X] \geq 0$.

On en déduit que si $X : \Omega \to \mathbb{R}$ est bornée, et $m \leq X(\omega) \leq M$ pour tout $\omega \in \Omega$, alors X est intégrable, et

$$m \leq E[X] \leq M. \tag{5.25}$$

On doit juste vérifier que X est intégrable, puis appliquer la majoration ci-dessus. Et puisque $|X(\omega)| \leq \max(|m|, |M|)$, et qu'on a vu que les fonctions constantes sont intégrables, on trouve bien que $E[|X|] \leq E[\max(|m|, |M|)] = \max(|m|, |M|)$.

Remarque. Si X est minorée (il existe $m \in \mathbb{R}$ tel que $X(\omega) \geq m$ pour $\omega \in \Omega$, on peut encore définir $E[X] \in \mathbb{R} \cup \{+\infty\}$, même si X n'est pas intégrable, en posant $E[X] = m' + E[X - m']$, où l'on a choisi n'importe quel $m' < m$; c'est parce que $X - m' \geq X - m \geq 0$. Mon conseil est de ne pas trop abuser de ceci, et de considérer directement $X - m$, ou $X - m'$, et de travailler avec.

Voilà. On est prêt pour de nouvelles lois !

5.6 Loi de Poisson, ou loi des événements rares.

En fait on va voir qu'il faudrait sans doute dire événements rares concernant de grandes populations, parce qu'il y aura une certaine compensation des deux.

Quelle est la bonne description pour la loi du nombre d'appels reçus en un jour par un dentiste qui a 3000 clients inépendants, qui téléphonent aléatoirement une fois tous les ans pour un rendez-vous ? On va expliquer que le bon modèle pour cette loi est une loi de Poisson (de Monsieur Siméon Denis Poisson, 1781-1840, apparament fait en 1838). Mais je commence par donner la définition, puis on verra comment cette loi est justifiable.

Définition 5.11. *Pour tout $\lambda > 0$, on définit une mesure de probabilité sur $\mathbb{N}$, appelée loi de Poisson de paramètre λ, à partir du germe de probabilité p_λ défini par*

$$p_\lambda(k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad \textit{pour } k \in \mathbb{N} \tag{5.26}$$

On note cette mesure Poisson(λ). *A cause de la formule qu'on verra plus loin, on peut aussi appeler dette mesure "loi de Poisson d'espérance λ".*

Ainsi, on dira qu'une variable aléatoire X (sur Ω, P) suit la loi de Poisson de paramètre $\lambda > 0$ si

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \text{ pour tout } k \in \mathbb{N}. \tag{5.27}$$

On a juste deux petites choses à vérifier. La première est que p_λ est bien un germe de probabilité, donc que $\sum_{k\geq 0} p_\lambda(k) = 1$. Ou encore, en multipliant par e^λ, que $\sum_{k\geq 0} \frac{\lambda^k}{k!} = e^\lambda$. J'espère que vous aurez eu le temps de voir que c'est en effet une des définitions de l'exponentielle.

La seconde vérification est que vous auriez pu croire que j'ai oublié de dire que X est à valeurs dans $\mathbb{N}$, ou au moins que $P(X \notin \mathbb{N}) = 0$. Mais en fait ceci se déduit du fait que si on a (5.26), alors par union disjointe $P(X \in \mathbb{N}) = \sum_{k\geq 0} P(X = k) = \sum_{k\geq 0} p_\lambda(k) = 1$, ce qui fait que $P(X \notin \mathbb{N}) = 0$ par passage au complémentaire.

Exemple. Maintenant, d'où sort cette formule bizarre ? Prenons l'exemple du dentiste, avec n clients (n est grand), mais qui se manifestent de manière aléatoire indépendante. Disons, chacun avec la probabilité λ/n de téléphoner chaque jour. On pense déjà que le nombre moyen d'appels par jours devrait être λ si on devine bien, d'où la normalisation choisie, mais on ne le vérifiera que plus tard.

Quelle est la loi du nombre X d'appels reçus un jour donné (tous les jours sont les mêmes) ? En fait on connaît cette loi, c'est la loi binomiale de paramètres n et p, où $p = \lambda/n$. En effet, chaque client donne une loi de Bernoulli de téléphoner, de paramètre p, et ensuite (4.17) dit que (si les clients agissent indépendamment) la loi de X, qui est la loi d'une somme de n Bernoulli paramètre p, est bien la loi binomiale de paramètres n et p.

Mais comme on a dit plus haut, quand n est grand on n'a pas trop envie de calculer les C_n^k, donc on cherche un équivalent de $P(X = k) = p^k(1-p)^{n-k}C_n^k$ (voir (4.16)). Vérifions que pour tout choix de k et λ,

$$\lim_{n\to+\infty} (\lambda/n)^k(1-(\lambda/n))^{n-k}C_n^k = e^{-\lambda}\frac{\lambda^k}{k!}; \tag{5.28}$$

le membre de gauche est la probabilité de $X = k$ pour la loi binomiale ci-dessus, puisque $p = \lambda/n$, et le membre de droite est la probabilité de $X = k$ pour la loi de Poisson. On verra ensuite comment traduire.

On approxime $C_n^k = \frac{n(n-1)\ldots(n-k+1)}{k!}$ par $\frac{n^k}{k!}$, au sens où le rapport tend vers 1.

Ensuite on garde $(\lambda/n)^k$. Le produit des deux tend vers $\frac{\lambda^k}{k!}$.

On peut sortir $\left(\frac{1}{1-(\lambda/n)}\right)^k$, qui tend vers 1 parce que λ/n tend vers 0 et k est fixe.

Il reste $(1-(\lambda/n))^n$, qui tend vers $e^{-\lambda}$. C'est une limite classique ; si vous ne la connaissez pas, prenez le logarithme : $\ln((1-(\lambda/n))^n = n\ln((1-(\lambda/n)) = n[-(\lambda/n) + o(1/n)]$, qui tend vers $-\lambda$; on reprend l'exponentielle qui est une fonction continue et on trouve le résultat.

Donc on a (5.28), qu'on va traduire sous forme de proposition, comme on a déjà fait dans le passé pour approximer la loi hypergéométrique par une loi binomiale (les proportions de poissons dans une grosse pêche, ou de gauchers dans un gros échantillon).

Proposition 5.12. *On se donne $\lambda > 0$, et une suite de variables aléatoires X_n. On suppose que X_n suit une loi binomiale de paramètres n et n/λ. Alors pour tout $k \geq 0$,*

$$\lim_{n\to+\infty} P(X_n = k) = e^{-\lambda}\frac{\lambda^k}{k!} \tag{5.29}$$

Donc en un sens, les lois des X_n tendent vers la loi de Poisson de paramètre λ. En fait, en regardant un peu mieux la démonstration, on aurait pu supposer que X_n suit une loi binomiale de paramètres N_n et p_n, à condition que $\lim_{n\to+\infty} N_n = +\infty$, et $\lim_{n\to+\infty} p_n N_n = \lambda$. C'est d'ailleurs rassurant, parce que prendre n/λ me rendait nerveux (ça nous oblige à prendre n/λ entier à chaque fois, ce qui donne des contraintes que ne devraient pas être).

Démonstration déjà faite, puisque $P(X_n = k)$ est le membre de gauche de (5.28). □

Encore un exemple. Il pleut des gouttes, et chaque goutte tombe de manière aléatoire sur une grande surface. On évalue qu'au cours d'une forte pluie il est tombé 10cm d'eau (c'est pas mal !) On estime qu'une goutte fait à peu près un vingtième de millilitre. Quelle est la loi du nombre de gouttes qui sont tombées sur une surface de 1mm^2 ? Corrrigez mes calculs si je me trompe mais je tombe sur une loi de Poisson d'espérance 2. En gros, 10 centimètres donnent $0.1m^3$

par m^2, donc 100 litres (oui, c'est assez lourd !), donc $2 \cdot 10^6$ gouttes. Donc 2 par mm^2. C'est bien sûr une très légère approximation.

Remarque. Il y a un équilibre entre la rareté du phénomène ($p = \lambda/n$ est petit) et le nombre n qui doit compenser. Encore que l'estimation ci-dessus est encore précise quand λ est tout petit (mais la probabilité de voir autre chose que 0 est alors faible).

Encore un exemple (on s'amuse bien). Il y a des poissons dans un lac, on me laisse entendre qu'il y en a 10^{-2} par m^2, disons répartis aléatoirement (ma meilleure évaluation ; de toute façon, la suite montre que je n'y connais rien). Je décide donc de pêcher à la grenade : j'en lance une et je tue tous les poissons qui sont à moins de 10m. Quelle est la loi du nombre de poissons récoltés ?

Plus sérieusement, la loi de Poisson est aussi ce qu'on utilise pour décrire le nombre d'atomes radioactifs dans un échantillon donné qui se décomposent en une minute, sachant la (petite) probabilité de chacun de se décomposer chaque minute, et le (grand) nombre d'atomes. Mais mon calcul rapide de ce matin semble montrer que si on regarde quelques grammes de produit radioactif et un temps d'une seconde, le nombre d'éléments qui se décomposent (en moyenne) est déjà très grand, de sorte que la loi de Poisson est avec un grand paramètre λ, et les fluctuations relatives de ce nombre seront très petites. Donc je ne suis pas certain que ce soit un si bon exemple, sauf à prendre des temps plus courts ou des nombres de particules bien plus petits que le nombre d'Avogadro qui est de l'ordre de $6 \cdot 10^{23}$.

Retour aux calcul. Comme promis, on peut calculer l'espérance d'une variable de Poisson.

Proposition 5.13. *Soit X une variable aléatoire qui suit une loi de Poisson de paramètre λ. Alors $E[X] = \lambda$.*

Comme X est à valeurs positives, pas même besoin de vérifier si X est intégrable, on verra bien à la fin si $E[X] < +\infty$. La formule donne

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} kP(X=k) = \sum_{k=1}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^k}{(k-1)!} \\ &= e^{-\lambda}\lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = e^{-\lambda}\lambda \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = e^{-\lambda}\lambda e^{\lambda} = \lambda \end{aligned}$$

où on a pu oter le terme où $k = 0$, puis on a posé $j = k-1$, et on s'est souvenir à la fin de la formule pour l'exponenielle. □

Il y a une certaine logique dans le cas de nos exemples ; s'il pleut en moyenne 2 gouttes par mm^2, on s'attend à ce que l'espérance de notre modèle soit bien 2. Autrement, on aurait probablement mal choisi notre modèle.

De la même manière, vérifions que la somme de deux variables de Poisson idépendante est une variable de Poisson (dont l'espérance est, naturellement, la somme des espérances).

Proposition 5.14. *Soient X et Y deux variables aléatoires indépendantes. On suppose que X suit une loi de Poisson de paramètre λ et que Y suit une loi de Poisson de paramètre μ Alors $X+Y$ suit une loi de Poisson de paramètre $\lambda+\mu$*

L'énoncé appelle plein de commentaires (je crois), mais commençons par la démonstration, qui doit mettre tout le monde d'accord. On se donne donc X et Y comme dans l'énoncé, et on calcule la loi de $X+Y$. Donc on calcule (5.30)

$$P(X+Y=k) = \sum_{j=0}^{k} P(X=j \text{ et } Y=k-j) = \sum_{j=0}^{k} P(X=j)P(Y=k-j)$$
$$= \sum_{j=0}^{k} \left(e^{-\lambda}\frac{\lambda^j}{j!}\right)\left(e^{-\mu}\frac{\mu^{k-j}}{(k-j)!}\right) = e^{-(\lambda+\mu)} \sum_{j=0}^{k} \frac{\lambda^j \mu^{k-j}}{j!(k-j)!}$$

où l'on a bien utilisé l'indépendance pour la seconde égalité, puis on a utilisé la loi, et enfin on a commencé à calculer. On remarque que $\frac{1}{j!(k-j)!} = \frac{C_k^j}{k!}$, donc

$$\sum_{j=0}^{k} \frac{\lambda^j \mu^{k-j}}{j!(k-j)!} = \frac{1}{k!}\sum_{j=0}^{k} C_k^j \lambda^j \mu^{k-j} = \frac{1}{k!}(\lambda+\mu)^k$$

(c'est bien pratique les coefficients binomiaux). On remplace dans (5.30) et on trouve que $P(X+Y=k) = e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^k}{k!}$, ce qui est exactement la probabilité de $\{k\}$ pour une loi de Poisson d'intensité $\lambda+\mu$. □

Passons aux commentaires.

Est-ce aussi miraculeux que les calculs semblent l'indiquer ? Je prétends que non. Reprenons l'exemple du dentiste, qui a n clients qui téléphonent en moyenne tous les λ/n jours. On aurait pu couper arbitrairement sa clientèle en deux, donc écrire $n = n_1 + n_2$, imaginer qu'il a deux clientèles, faire le calcul pour chacune avec sa loi de Poisson, et obtenir le résultat prévu.

Dit de manière plus constructive, imaginons que le dentiste recoive un jour la clientèle d'un collègue qui part à la retraite. Il avait n_1 clients, qui téléphonaient avec la probabilité $p = \lambda/n_1$ tous les jours, et maintenant il se retrouve avec $n_1 + n_2$ clients semblables, qui téléphonent avec la même probabilité p. On a dit que chaque jour la loi du nombre d'appels de sa première clientèle est (en gros) une loi de Poisson d'espérance $\lambda = pn_1$. Pour son collègue, c'était une loi de Poisson d'espérance $\mu = pn_2$. Et pour la somme des deux, c'est normal que ce soit une loi de Poisson d'espérance $p(n_1 + n_2) = \lambda + \mu$. Si ça n'avait pas été le cas, le modèle aurait été plutôt mauvais.

Ceci ne donne pas une démonstration, parce que notre modèle n'est pas exact. Dans le cas des lois binomiales, ceci donnerait une démonstration, celle qu'on a faite au corollaire 4.12. Ensuite, si on voulait en déduire la proposition 5.14, il faudrait encore se fendre d'un argument d'approximation un peu délicat. Ce n'est pas utile, puisqu'on a une démonstration simple.

Est-ce à dire que la proposition 5.14 est inutile ? Pas du tout, mais alors là, pas du tout ! Reprenons notre dentiste. Il s'aperçoit avec surprise que les clients de son collègue téléphonent en fait avec une probabilité $p_2 \neq p$. Donc tous ses calculs s'effondrent, et il commence même à avoir des doutes sur la manière dont il a modélisé toute son histoire. Parce qu'au fond, lui aussi a divers types de clients, avec des habitudes différentes, alors notre modèle initial était sans doute naif et faux ? Heureusement, à cause de la proposition, pas tant que ça.

Reprenons le modèle avec juste deux types de clients ayant des comportement différents. Les premiers appels ont une loi de Poisson d'espérance λ, les second, ont une loi de Poisson d'espérance μ, on suppose que les deux populations téléphonent de manière indépendante, et finalement la proposition dit que la loi de la somme est encore une loi de Poisson. Ce qu'il faut réviser, c'est juste la formule qui donne l'espérance (en fonction de n et de p), pour chacune des populations, mais c'est quand même beaucoup moins grave.

Et surtout ceci renforce notre idée que la loi de Poisson donne un bon modèle

pour des sommes d'événements rares, puisqu'on peut mélanger des populations inhomogènes et quand même obtenir une loi de Poisson !

Un autre exemple, plus d'actualité (et que je permets en dépit du fait que nous avons vu quelques exemples de cas où un scientifique peut dire des bêtises quand il sort de son domaine de compétence ; vous me corrigerez si je déraille trop). On veut modéliser la probabilité du nombre de personnes qui entrent en réanimation chaque jour, par exemple en rapport avec la covid. A cause de la discussion ci-dessus, on se doute bien que le modèle va être une loi de Poisson. Mais avec quelle espérance ? Si on veut l'estimer, le plus raisonnable est de couper les populations en catégories plus ou moins précises (en fonction du risque et de l'age, voire de l'adresse et tout ce qu'on sait), calculer l'espérancer de la loi de Poisson correspondante, et juste additionner les espérances. C'est drôlement pratique ! Même si bien sûr pour calculer toutes ces espérances, il est bon aussi d'avoir une idée du nombre de malades, de la fréquence p à laquelle leur état se dégrade, etc.

Si je dirige un hôpital, que j'ai 4 places en réanimation, et qu'on me dit que les malades arrivent chez moi avec une espérance 1 chaque jour, je peux calculer la probabilité d'avoir au moins 5 arrivées dans les deux prochains jours : c'est $P(X \geq 5)$ pour une loi de Poisson d'espérance 2, et c'est très facile à calculer. Si cette probabilité est trop grande, je peux commencer à m'inquiéter d'un point de chute pour les suivants.

Un dernier commentaire : dans la proposition on suppose que les variables sont indépendantes. Sinon, on ne peut pas faire grand-chose comme calculs, mais ne jamais oublier que c'est une hypothèse forte, rarement réalisée de manière exacte dans la pratique. Dans l'exemple du dentiste, si la veille un reportage alarmiste sur le sucre est passé à la télé, il est probable que la loi des appels sera encore une loi de Poisson, mais pas avec le même λ. Et si dans les clients il y a des groupes d'amis qui font toujours la même chose le même jour, il faudrait en tenir compte et Poisson s'effondre.

6 Inégalité de Markov, variance, loi des grands nombres

On va faire quelques estimations sur les variables aléatoires (à valeurs réelles). On se dirige vers la loi des grands nombres, mais le trajet est intéressant aussi.

6.1 Inégalité de Markov

Je commence par une inégalité simple, qui en probas s'appelle l'inégalité de Markov et en analyse l'inégalité de Chebyshev. Je la donne d'abord pour une v.a. à valeurs positives, puis on passera facilement au cas réel.

Proposition 6.1. *Soient $X : \Omega \to [0, +\infty)$ une variable aléatoire, et $\lambda > 0$ un nombre strictement positif. Alors*

$$P(\{X(\omega) \geq \lambda\}) \leq \frac{E[X]}{\lambda}. \tag{6.1}$$

La prochaine fois j'écris directement $P(X \leq \lambda)$. C'est vrai, mais totalement inutile, si $E[X] = +\infty$, donc on aurait aussi pu supposer $E[X] < +\infty$. Pareil si λ est si petit que le membre de droite est > 1, d'ailleurs.

Et ceci est aussi un résultat utile valable pour l'intégrale (de Riemann, et bientôt de Lebesgue) des fonctions.

C'est très utile, mais la démonstration est simple. Notons $A = \big\{\omega \in \Omega\,;\, X(\omega) \geq \lambda\big\}$. On veut montrer que $P(A) \leq \frac{E[X]}{\lambda}$. Mais la variable aléatoire Y qui vaut λ sur A et 0 hors de A est inférieure à X, donc $E[Y] \leq E[X]$. Or $E[Y]$ est facile à calculer, puisque cette fonction prend juste deux valeurs, et

$$E[Y] = \lambda P(Y = \lambda) + 0P(Y = 0) = \lambda P(Y = \lambda) = \lambda P(A).$$

Donc finalement $P(A) = \lambda^{-1}E[Y] \leq \lambda^{-1}E[X]$ comme annoncé. □

Je me permets quand même de donner la version de ceci pour les intégrales : si f est une fonction (intégrable) positive sur un intervalle I, cette inégalité s'écrit

$$\big|\big\{x \in I\,;\, f(x) > \lambda\big\}\big| \leq \lambda^{-1} \int_I f(x)dx,$$

où $|A| = \int_I \mathbb{1}_A(x)dx$ est la mesure de l'ensemble x. Et la démonstration consiste à dire que si on pose $A = \big\{x \in I\,;\, f(x) \geq \lambda\big\}$ comme ci-dessus, alors

$$\lambda|A| = \lambda \int_I \mathbb{1}_A(x)dx = \int_I \lambda\mathbb{1}_A(x)dx \leq \int_I f(x)dx$$

à nouveau parce que $f \geq \lambda\mathbb{1}_A$, ou on écrit de manière un peu différente

$$\lambda|A| = \int_A \lambda dx \leq \int_A f(x)dx \leq \int_I f(x)dx.$$

Revenons à Markov pour les mesures de probabilité. Voici l'énoncé pour les fonctions à valeurs réelles.

Proposition 6.2. *Soient $X : \Omega \to \mathbb{R}$ une variable aléatoire intégrable (telle que $E[|X|] < +\infty$), et $\lambda > 0$ un nombre strictement positif. Alors*

$$P(\{|X(\omega)| \geq \lambda\}) \leq \frac{E[|X|]}{\lambda}. \tag{6.2}$$

Vous voyez que c'est en fait la proposition précédente, appliquée à la v.a. $|X|$. Et d'ailleurs si X n'est pas intégrable, $E[|X|] = +\infty$, et le résultat est vrai mais inutile. □

Applications ci-dessous, dans le cas de la variance.

Mais en gros, si vous avez calculé que votre espérance de gain à un jeu est 1, vous savez que votre probabilité de gagner 100 est inférieure à 10^{-2}. Elle peut naturellement être strictement plus petite, par exemple si le jeu n'autorise pas de tels gains.

Pareillement, si la variable X suit une loi de Poisson de paramètre $\lambda = 2$, donc son espérance est 2, le théorème dit que la probabilité $P(X \geq 6)$ est inférieure à $2/6 = 1/3$. C'est bien parce que ce serait vrai avec n'importe quelle v.a. ayant cette espérance, mais si je sais que c'est une loi de Poisson, je peux utiliser la formule (5.27). Mex calculs rapides mais peut-être faux donnent $P(X \geq 6) < 0,02$.

6.2 Variance

C'est le prochain nombre intéressant après l'espérance.

Définition 6.3. *Soient $X : \Omega \to \mathbb{R}$ une variable aléatoire intégrable (telle que $E[|X|] < +\infty$), La variance de X, souvent notée $Var(X)$, est le nombre positif*

$$Var(X) = E[(X - E(X))^2]. \tag{6.3}$$

On appelle écart type de X le nombre $\sigma(X) = \sqrt{Var(X)}$.

Fin du cours 9, 2020

Des commentaires d'abord.

On dit que la v.a. X est centrée lorsque $E[X] = 0$. Souvent on préfère travailler sur les v.a. centrées, et c'est facile puisque si X est une v.a. d'espérance finie (intégrable si vous préférez), on retire sa moyenne et on trouve une nouvelle variable $X - E[X]$ qui est centrée !

Pour les variables centrées, la formule est plus simple, puisqu'alors $Var(X) = E[X^2]$.

Rien ne dit que la variance est finie, même si X est intégrable. Par exemple, pour la mesure de probabilité P sur $\mathbb{N}$ telle que $P(\{n\}) = 2^{-n-1}$, la variable X donnée par $X(n) = 2^{n/2}$ est intégrable, mais pas son carré.

Quand on ajoute une constante à X, on ne modifie ni sa variance ni son écart type. C'est fait exprès.

Quand on multiplie X par λ, on multiplie $Var(X)$ par λ^2 et $\sigma(X)$ seulement par $|\lambda|$. L'idée est que l'écart type donne une idée de la distance moyenne (mais c'est mieux de dire type) entre X et sa moyenne. Plus les valeurs de X sont concentrées autour de $E[X]$, plus l'écart type est petit. Mais ce qui vaut, bien sûr, c'est la définition !

Malgré le fait que l'homogénéité de σ est la même que celle de X, c'est souvent la variance qui tombe dans les calculs. Par contre, pour une variable centrée, c'est bien $\sigma(X)$ qui donne une idée de sa taille habituelle.

Vous allez dire, pourquoi on n'a pas choisi de travailler avec $E[|X - E[X]|]$ qui a directement la bonne homogénéité ? C'est à cause de la magie des carrés (ou des produis scalaires) ; on verra que les calculs s'arrangent mieux avec la variance.

Comme d'habitude, $Var(X)$ et $\sigma(X)$ ne dépendent que de la loi de X (tout comme l'espérance).

Pour moi : faire un dessin d'histogramme de la loi de X et demander qui donne le plus grand écart type.

Des propriétés élémentaires (mais utile) ensuite. Je regroupe même si on en a vu certaines.

Proposition 6.4. *Soient $X : \Omega \to \mathbb{R}$ une variable aléatoire intégrable (telle que $E[|X|] < +\infty$), et $\lambda > 0$ un nombre strictement positif. Alors*
(6.4)
$Var(x) = 0$ si et seulement si X est constante, égale presque partout à $E[X]$;

(6.5) $$Var(\lambda X) = \lambda^2 Var(X) \quad \textit{pour tout } \lambda > 0\,;$$

(6.6) $$Var(X) = E[X^2] - E[X]^2$$

Noter qu'au passage (6.6) dit que $E[X^2] \geq E[X]^2$ (justement à cause de la variance !).

Démonstration. On a vu que $Var(x) \geq 0$ (espérance de la v.a. $(X - E[X])^2$ qui est positive. Il est clair que $Var(x) = 0$ quand X est constante (donc égale à $E[X]$). Il reste la réciproque. J'ai ajouté presque-partout parce qu'en fait on ne va démontrer que le fait que $P(X \neq E[X]) = 0$. Il se pourrait qu'on ait pris $P(X) \neq E[X]$ sur un ensemble non vide, mais de probabilité nulle. Probablement vous avez envie de retirer cet ensemble de probabilité nulle de Ω, pusiqu'il ne sert à rien, mais en attendant nos hypothèses ne nous permettent pas de l'éliminer a priori. Bon, on note p le germe de P, ent on note que

$$Var(X) = E[(X - E[X])^2] = \sum_{\omega \in \Omega} (X(\omega) - E[X])^2 p(\omega).$$

C'est une somme de termes positifs. Si la somme est nulle, tous les termes sont nuls. Donc pour tout ω, ou bien $X(\omega) = E[X]$ comme on a annoncé, ou alors $p(\omega) = 0$ (et on retrouve l'événement de probabilité nulle où X pourrait être différent, mais ça n'est pas grave puisque cet événement ne se produit pas).

Passons à (6.5) (annoncé plus haut). On fait juste le calcul en notant que $E[X]$ est aussi multiplié par λ.

Pour (6.6) c'est vraiment plus intéressant (la moitié du temps on calcule $Var(X)$ comme ceci). Notons $e = E[X]$ pour simplifier les notations. Ainsi $(X - E[X])^2 = (X - e)^2 = X^2 - 2eX + e^2$; son espérance est $Var(X)$ par définition, et d'autre part vaut $E[X^2] - 2eE[X] + e^2 = E[X^2] - 2e^2 + e^2 = E[X^2] - e^2$. On en déduit (6.6). □

Il pourra être utile de savoir que variance finie implique espérance finie. Plus précisément, on dit que la variable X est de carré intégrable, ou de variance finie, quand $E[X^2] < +\infty$ (noter que $E[X^2]$ est toujours défini, puisque $X^2 \geq 0$, mais pourrait valoir $+\infty$). Vérifions que

(6.7) si X est de variance finie, alors elle est intégrable.

En effet, il est vrai que $|2x| \leq x^2 + 1$ pour tout $x \in \mathbb{R}$ (juste parce que $(|x| - 1)^2 \geq 0$, en développant). Donc $|X(\omega)| \leq (1 + X(\omega)^2)/2$ pour tout ω, et en prenant les espérances, $E[|X|] \leq (E[X^2] + 1)/2 < +\infty$ si X est de variance finie.

6.3 Variance et indépendance

Proposition 6.5. *Soient $X, Y : \Omega \to \mathbb{R}$ deux variable aléatoire intégrables (telles que $E[|X|] < +\infty$ et $E[|Y|] < +\infty$) indépendantes. Alors XY est intégrable et*

$$E[XY] = E[X]E[Y]. \tag{6.8}$$

Evidemment c'est faux en général quand X et Y ne sont pas indépendantes ; par exemple on a vu que $E[X^2] > E[X]^2$ dès que X n'est pas constante, puisque la différence est la variance de X.

On commence par démontrer ceci quand X et Y sont (à valeurs) positives. On part de la formule

$$E[XY] = \sum_{\omega \in \Omega} p(\omega) X(\omega) Y(\omega),$$

où p est le germe de P, et où a priori on ne sait pas encore si la somme est finie ou pas. N'empêche qu'on peut prendre la somme de droite, et la découper à l'aide d'une partition, en fonction des valeurs de $X(\omega)$.

$$E[XY] = \sum_{x \in X(\omega)} \Big\{ \sum_{\omega \in \Omega ; X(\omega) = x} p(\omega) X(\omega) Y(\omega) \Big\}$$

où le théorème sur les sommes avec des partitions dit qu'il y a égalité même si l'un des deux membres est infini. On recommence en partitionnant chacune des sommes intérieures par les valeurs de $Y(\omega)$. On trouve

$$\begin{aligned} E[XY] &= \sum_{x \in X(\omega)} \Big\{ \sum_{y \in Y(\omega)} \Big\{ \sum_{\omega \in \Omega ; X(\omega) = x \text{ et } Y(\omega) = y} p(\omega) X(\omega) Y(\omega) \Big\} \Big\} \\ &= \sum_{x \in X(\omega)} \Big\{ \sum_{y \in Y(\omega)} \Big\{ \sum_{\omega \in \Omega ; X(\omega) = x \text{ et } Y(\omega) = y} xyp(\omega) \Big\} \Big\} \\ &= \sum_{x \in X(\omega)} \Big\{ \sum_{y \in Y(\omega)} xy P\big(X(\omega) = x \text{ et } Y(\omega) = y\big) \Big\}. \end{aligned}$$

Mais X et Y sont indépendantes, donc $P\big(X(\omega) = x \text{ et } Y(\omega) = y\big) = P(X(\omega) = x)P(Y(\omega) = y)$. On remplace dans la somme et on trouve que

$$E[XY] = \sum_{x\in X(\omega)} \Big\{ \sum_{y\in Y(\omega)} xyP(X = x)P(Y = y)\Big\}$$
$$= \sum_{x\in X(\omega)} xP(X = x) \Big\{ \sum_{y\in Y(\omega)} yP(Y = y)\Big\} = \sum_{x\in X(\omega)} xP(X = x)E[Y] = E[X]E[Y].$$

Quand on suit la démonstration, on voit en particulier qu'avec notre hypothèse que $E[X] < +\infty$ et $E[X] < +\infty$, on trouve $E[XY] < +\infty$.

On passe au cas des variables intégrébles à valeurs réelles. Noter que $|X|$ et $|Y|$ sont encore indépendantes, en vertu de la proposition 4.8 (ou alors recalculer $P(|X| = x \text{ et } |Y| = y)$ à partir d'expressions semblables pour X et Y, donc par le cas qu'on vient de faire $E[|XY|] = E[|X|]E[|Y|] < +\infty$.

Maintenant on peut calculer $E[XY]$ comme ci-dessus, avec des sommes de familles sommables, appliquer le résultat sur les sommations par paquets, réécrire les calculs ci-dessus mot pour mot, et obtenir le résultat souhaité (6.8). □

Théorème 6.6. *Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires indépendantes à valeurs réelles. Supposons en outre que chaque X_j est de variance finie (on dit aussi de carré intégrable), c.-à-d. que $E[|X_j|^2] < +\infty$. Alors $X_1 + X_2 + \ldots X_n$ est aussi de variance finie, et*

$$Var(X_1 + X_2 + \ldots X_n) = Var(X_1) + Var(X_2) + \ldots Var(X_n). \tag{6.9}$$

Ca marcherait d'ailleurs aussi pour des v.a. indépendantes à valeurs dans $\mathbb{C}$, ou dans $\mathbb{R}^n$, au lieu de $\mathbb{R}$.

On a vu que les X_j sont automatiquement d'espérance finie, parce qu'elles sont de variance finie (voir (6.7)).

On commence par le cas où $n = 2$. Comme je me perds aisément dans les notations, Notons $e = E[X_1]$ et $f = E[X_2]$ pour essayer de clarifier. Le fait que $X_1 + X_2$ est intégrable est une conséquence déjà vue du fait que X_1 et X_2 sont d'espérances finies. Le fait que $X_1 + X_2$ a une variance finie se déduirait du calcul ci-dessous, mais vérifions-le à l'avance pour se rassurer ; de toute manière

c'est une majoration standard utile. On observe simplement que pour tout ω, $(X_1(\omega)+X_2(\omega))^2 = X_1(\omega)^2 + X_2(\omega) + 2X_1(\omega)X_2(\omega) \leq 2X_1(\omega)^2 + 2X_2(\omega)^2$ parce qu'il est connu que $|ab| \leq a^2+b^2$ pou $a, b \in \mathbb{R}$, simplement parce que $(a-b)^2 \geq 0$. La conséquence est que la variable aléatoire $(X+Y)^2$ est dominée par $2X^2+2Y^2$, qui est d'espérance finie par hypothèse.

Revenons au calcul. Alors

(6.10)
$$Var(X_1+X_2) = E\big[(X_1+X_2-E[X_1+X_2])^2\big] = E\big[(X_1+X_2-e-f)^2\big]$$
$$= E\big[((X_1-e)+(X_2-f))^2\big] = E\big[(X_1-e)^2+2(X_1-e)(X_2-f)+(X_2-f)^2\big]$$
$$= E[(X_1-e)^2]+E[(X_2-f)^2]+2E[(X_1-e)(X_2-f)] = Var(X_1)+Var(X_2)+2E[(X_1-e)(X_2-$$

On a juste calculé et utilisé les définitions. Maintenant on utilise l'indépendance. Si X_1 et X_2 sont indépendantes, c'est aussi vrai de X_1-e et X_2-f, qui sont une fonction de X_1 et une fonction de X_2 (utiliser la proposition 4.8 ou directement la définition d'indépendance). Donc $E[(X_1-e)(X_2-f)] = E[(X_1-e)]E[(X_2-f)] = 0$ parce que $E[X_1-e] = E[X_1]-e = 0$ par définition de e, et d'ailleurs pareil pour $E[(X_2-f)]$. Il reste $Var(X_1+X_2) = Var(X_1)+Var(X_2)$ comme souhaité.

Dans le cas de n variables, on démontre (6.9) par récurrence sur n, en remarquant pour commencer que si les variables $X_1, \ldots, X_n$ sont (globalement !) indépendantes, alors en particulier X_1 est indépendante de $Y = X_2+\ldots+X_n$. On applique la formule pour la somme $X+Y$, et par hypothèse de récurrence la formule pour $X_2+\ldots+X_n$, et on obtient le résultat souhaité. □

Le résultat est très pratique, et d'ailleurs on s'en servira bientôt. Sans supposer l'indépendance, dans le calcul de $Var(X_1+X_2)$ on a un terme supplémentaire $2E[(X_1-e)(X_2-f)]$, donc le résultat (pour $n=2$) est faux en général. Alors on fait comme souvent une définition qui correspond, celle de la covariance de deux v.a.

Définition 6.7. *Soient X, Y deux variables aléatoires à valeurs réelles. On suppose que $E[|X|] < +\infty$, $E[|Y|] < +\infty$, et $E[|XY|] < +\infty$. Alors on appelle covariance de X et Y le nombre*

(6.11) $$Cov(X,Y) = E\big[(X-E[X])(Y-E[Y])\big] = E[XY]-E[X]E[Y].$$

Comme d'habitude il y a une petite vérification à faire. Le membre de droite est défini directement à partir des hypothèses. Pour le membre de gauche, notons encore $e = E[X]$ et $f = E[Y]$ pour y voir plus clair ; alors $(X - E[X])(Y - E[Y]) = XY - eY - fX + ef$; on en déduit bien que le membre de gauche a un sens (somme de variables d'espérances finies). Et quand on calcule l'espérance de $(X - E[X])(Y - E[Y])$, on trouve bien $E[XY] - eE[Y] - fE[X] - ef = E[XY] - ef - ef + ef = E[XY] - ef$ comme annoncé.

Les propriétés simples de la covariances sont les suivantes. D'abord

$$Cov(X, X) = Var(X) \tag{6.12}$$

(quand X a une variance finie, donc aussi une espérance finie) ;

$$Cov(X, Y) = 0 \text{ quand } X \text{ et } Y \text{ sont indépendantes} \tag{6.13}$$

(et en supposant que les deux ont une espérance finie), à cause de la proposition 6.5. Enfin, toujours pour des variables X et Y qui ont une variance finie (donc aussi une espérance finie),

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y). \tag{6.14}$$

Ceci, à cause du calcul qui donne (6.10), qui dit directement que XY est d'espérance finie (donc on peut calculer) et que $Var(X + Y) = Var(X) + Var(Y) + 2E\big[(X - E[X])(Y - E[Y])\big]$, voir (6.10).

6.4 Variance de lois classiques

Pour ne pas avoir à les recalculer à chaque fois. Certaines, sans démonstration. On rapelle que la variance, tout comme l'espérance ne dépend que de la loi de la variable.

Pas de problème d'existence pour Ω fini : la variance est forcément finie !

Proposition 6.8. *La variance d'une (v.a. de) loi de Bernoulli de paramètre p est $p(1 - p)$.*
La variance d'une loi binomiale de pamètres n et p est $np(1 - p)$.

Démonstration. Pour Bernoulli (pile ou face avec une pièce lestée), on peut faire la calcul : $E[X] = p$ (une seule contribution de $x = 1$). Puis $E[X^2] = p$

aussi, cette fois comme $0^2p(0)+1^2p(1)=p$. Et finalement $Var(X)=E[X^2]-E[X]^2=p-p^2$.

Pour la loi binomiale (typiquement, somme de n lancers indépendants de pièces lestées), on choisit un X particulier dont la loi est binomiale : on construit d'abord n lois de Bernoulli X_j de paramètre p et qui sont indépendantes. Par exemple en prenant $\Omega=E_2^n$ avec une probabilité produit. Puis on se souvient qu'alors la loi de $X=X_1+\dots X_n$ est une loi binomiale de paramèters n et p. Et enfin on applique le théorème 6.6 pour trouver que $Var(X)=\sum_{j=1}^n Var(X_j)=np(p-1)$ comme annoncé. Et bien sûr le résultat aurait été le même pour tout autre X binomial. C'est quand même bien pratique cette histoire ! □

Proposition 6.9. *La variance d'une (x.a. de) loi hypergéométrique de paramètres N, G, n est inférieure à $n\,\frac{G}{N}\,(1-\frac{G}{N})$.*

Sans démonstration mais avec commentaires. On se souvient que cette loi $\mathcal{H}(N,G,p)$ correspond à celle de n tirages sans remise dans une population de N personnes et avec une sous-population G de gauchers. Et plus précisément que pour $0\leq k\leq n$, $P(X=k)=\frac{C_G^k C_{N-G}^{n-k}}{C_N^n}$ est la probabilité de tirer k gauchers (parmi les G).

La loi se rapproche un peu (en particulier quend N et G sont grands par rapport à n) de la loi binomiale (n tirages sans remise, avec les chances égales à G/N et $1-G/N$ de tirer chaque population), et donc ça a un sens de comparer $Var(\mathcal{H}(N,G,p))$ avec la variance de la loi binomiale correspondante. Qui est bien $n\,\frac{G}{N}\,(1-\frac{G}{N})$.

Donc le résultat qu'on ne démontre pas est que pour la loi hypergéométrique, la variance est moindre. C'est un peu logique. On compare tirage avec remise et tirage sans remise, et la variance mesure une moyenne d'écarts (au carré) entre le tirage et sa valeur moyenne. Supposons qu'on ait à un moment tiré trop de gauchers. Avec remise, on continue au hasard. Sans remise, on a maintenant moins de chance de tirer des gauchers, puisqu'il en reste moins que dans le cas avec remise. Du coup on a un effet stabilisateur qui nous ramène un peu vers la moyenne et donc diminue un peu la variance. Plus G et N sont grands par rapport à n, plus cet effet est petit, ce qui est en accord avec le fait que la loi $\mathcal{H}$ ressemble de plus en plus à une loi binomiale.

Fin du cours 10, 2020

Proposition 6.10. *La variance d'une (x.a. de) loi géométrique de paramètre* $p \in]0,1]$ *est* $\frac{1-p}{p^2}$.

Sans démonstration. On se souvient que c'est la loi, sur $\mathbb{N}^*$, du premier lancement réussi de pièce lestée, et que $P(\{k\}) = p(1-p)^{k-1}$. Voir la définition 4.13. Pas trop surprenant que ce soit grand quand p est petit, parce les valeurs de X tendent à être grandes (il faut attendre longtemps pour gagner). De ce point de vue, l'écart type $\sqrt{\frac{1-p}{p^2}}$ est du même ordre de grandeur que $E[X] = \frac{1}{p}$.

Proposition 6.11. *La variance d'une loi de Poisson d'espérance* λ *est* λ.

Sans démonstration. Mais au moins il est logique que ce soit proportionnel à λ, puisque la somme de variables de Poisson X et Y de paramètres λ et μ est une variable de Poisson de paramètre $\lambda + \mu$, donc si on note $v(\lambda)$ la variance d'une loi de Poisson d'espérance λ, le théorème 6.6 dit que $v(\lambda + \mu) = v(\lambda) + v(\mu)$. En gros, il n'y a vraiment que la constante multiplicative 1 à calculer.

On a aussi vu que la loi de Poisson ressemblait beaucoup, quand n devient grand, à la loi binomiale de paramètres n et λ/n, dont on vient de voir que la variance est $n(\lambda/n)(1-\lambda/n)$ qui tend vers λ, donc au moins on n'est pas trop surpris (ni par la constante 1).

6.5 Inégalité de Bienaymé-Chebyshev

De Irénée-Jules Bienaymé (1796-1878) et Pafnouti Chebyshev (ou autres orthographes), 1821-1894. Une inégalité qui permet de prédire quelles sont les chances de s'éloigner de l'espérances, en fonction de la variance.

Théorème 6.12. *Soit* X *une variable aléatoire de variance finie. Alors pour tout* $\varepsilon > 0$,

$$P(|X - E[X]| \geq \varepsilon) \leq \frac{Var(X)}{\varepsilon^2}. \tag{6.15}$$

Démonstration (facile) plus bas ; on commence par des commentaires.

On a appelé ε le nombre strictement positif parce que parfois, dans les applications, il est petit (ou petit par rapport à l'écart type). Mais l'inégalité vaut pour tout $\varepsilon > 0$.

Une autre manière d'écrire (6.15), est de dire que pour tout $\lambda > 0$,

$$P\big(|X - E[X]| \geq \lambda\sigma(X)\big) \leq \frac{1}{\lambda^2}. \tag{6.16}$$

L'avantage de cette seconde écriture est qu'on voit bien que c'est le ratio entre l'écart $|X - E[X]|$ et l'écart type $\sigma(X) = \sqrt{Var(X)}$ qui donne une idée de la probabilité de l'événement. Vérifions que (6.15) implique (6.16). Soit $\lambda > 0$, et appliquons (6.15) avec $\varepsilon = \lambda\sigma(X)$. Alors

$$P\big(|X - E[X]| \geq \lambda\sigma(X)\big) = P\big(|X - E[X]| \geq \varepsilon\big) \leq \frac{Var(X)}{\varepsilon^2} = \frac{1}{\lambda^2}$$

puisque $Var(X) = \sigma(X)^2$. On passerait de même de (6.16) à (6.15), mais c'est moins important puisque c'est (6.15) qu'on va démontrer.

Le théorème est une inégalité. On ne peut pas faire mieux en général. Par exemple, prendre X à valeurs dans $\{-1, 1\}$ telle que $P(X = 1) = P(X = -1) = 1/2$. L'espérance est 0, la variance 1, et quand on prend $\varepsilon = 1$ dans (6.15) le membre de gauche est 1 (puisque $|X - E[X]| = |X|$ est toujours 1). Mais évidemment la plupart du temps l'inégalité de (6.15) est stricte. Quand même, (6.15) donne souvent une bonne idée de ce qui se passe (en gros, à un facteur près).

L'inégalité (6.15) est encore vraie quand on suppose seulement que X est d'espérance finie. Mais si la variance est infinie, le second membre aussi, et le résultat n'a aucun intérêt.

Il est temps de passer à la démonstration. On applique juste l'inégalité de Markov à la variable $Y = |X - E[X]|^2$. Alors

$$P(|X - E[X]| \geq \varepsilon) = P(Y \geq \varepsilon^2) \leq \frac{E[Y]}{\varepsilon^2} = \frac{E[|X - E[X]|^2]}{\varepsilon^2} = \frac{Var(X)}{\varepsilon^2}$$

par définition de la variance. □

Exemple typique d'application. On lance 20 fois une pièce équilibrée, et on se demande la probabilité d'avoir plus fait pile 15 fois ou plus. [Une idée, avant de faire le calcul ?]

Donc on note X la v.a. qui donne le nombre de piles, on se souvient que comme somme de Bernoulli indépendantes, X suit la loi binomiale de paramètres 20 et $1/2$, puis que l'espérance est 10 et que la variance est 20 fois $p(1-p)$, soit $20/4 = 5$.

Maintenant on applique BC, avec $\varepsilon = 5$ (pas si petit) et on trouve que la probabilité cherchée est

$$P(X \geq 15) \leq P(|X - 10| \geq 5) \leq \frac{Var(X))}{5^2} = \frac{5}{25} = \frac{1}{5}.$$

Bon en fait il est clair (par symétrie, vous êtes d'accord ?) que $P(X \geq 15) = P(X \leq 5)$, donc la probabilité cherchée est seulement la moitié de $P(|X-10| \geq 5)$, et est donc inférieure à $1/10$. Un exercice intéressant serait de voir, en faisant additionner les vraies probabilités d'avoir $15, 16, \ldots 0$, si on trouverait beaucoup mieux. On dirait que les tables statistiques disponibles dans WIMS disent environ $2/100$, ce qui est bien mieux.

6.6 Loi (faible) des grands nombres

Faible, parce qu'il y a des énoncés meilleurs et plus précis, mais qu'on ne verra pas.

Idée de base. On nous a donné une pièce lestée, et on veut savoir la probabilité p de tomber sur pile. On lance donc la pièce n fois, n assez grand, et on se dit que le meilleur moyen d'estimer p est de diviser par n le nombre de piles obtenus. On obtient un nombre $\widehat{p}$. C'est sans doute le mieux qu'on puisse faire (au moins sans analyser la pièce).

La question est de savoir de combien on risque de s'être trompé. On ne peut pas exclure que la pièce soit équilibrée, mais qu'on ait eu la malchance pendant notre essai de tomber toujours sur pile (et donc d'avoir une estimation fausse). Mais on sait bien aussi que le risque est faible, et en fait on veut l'évaluer précisément. En tout cas, le mieux qu'on puisse faire à ce stade est d'évaluer la probabilité que de se tromper de ε, c.-à-d le nombre

$$P(|\widehat{p} - p| > \varepsilon).$$

Cette question ressemble beaucoup à ce qu'on vient de faire dans l'exemple typique, mais on va quand même trouver utile de formaliser un peu plus tout ceci. Et surtout, d'obtenir des résultats même quand la loi du lancer devient plus compliquée qu'une loi de Bernoulli où on peut tout calculer.

Le problème semble un peu différent au départ : avant, on connaissait p et on voulait calculer la probabilité de certains lancers, alors que maintenant on

ne connait pas p, on a fait les lancers, et on veut savoir la probabilité de s'être beaucoup trompé. Mais la modélisation restera la même en gros : de toute façon on va considérer que les lancers ont une certaine loi, et on va calculer à partir de ça.

Dans l'énoncé qui suit on se donne une suite de variables indépendantes X_n. Ne vous demandez pas ce que c'est, c'est juste une suite de variables telles que pour tout $N \geq 1$, les variables $X_1, \dots X_N$ sont indépendantes. Et cette dernière chose signifie que pour tout choix de $x_1 \in X_1(\Omega)$, $x_2 \in X_2(\Omega), \dots x_N \in X_N(\Omega)$,

$$P\big(X_1 = x_1 \text{ et } X_2 = x_2 \dots \text{ et } X_N = x_N\big) = \prod_{n=1}^{N} P(X_n = x_n). \tag{6.17}$$

En fait on peut même le demander quand $x_n \notin X_n(\Omega)$, mais alors les deux membres de (6.17) sont nuls. Et on peut déduire de ceci, en faisant des regroupements de sommes, que pour tous choix d'ensembles A_1, $A_2, \dots A_N$,

$$P\big(X_1 \in A_1 \text{ et } X_2 \in A_2 \dots \text{ et } X_N \in A_N\big) = \prod_{n=1}^{N} P(X_n \in A_n). \tag{6.18}$$

Théorème 6.13. *Soit $\{X_j\}$, $j \in \mathbb{N}^*$, une suite de variables indépendantes et de même loi. On suppose aussi que $E[X_1^2] < +\infty$. On pose $S_n = \sum_{j=1}^{n} X_j$ pour $n \geq 1$. Alors, pour tout $\varepsilon > 0$,*

$$\lim_{n \to +\infty} P\Big(\Big|\frac{S_n}{n} - E[X_1]\Big| \geq \varepsilon\Big) = 0. \tag{6.19}$$

On appelle ceci la loi faible des grands nombres. On verra qu'on a un résultat plus précis, à savoir (6.20). Et aussi on pourrait faire des énoncés plus généraux. En fait c'est la démontration qui est importante.

Il s'agit donc d'une manière précise de dire que les moyennes $\frac{S_n}{n}$ convergent, quand n tend vers $+\infty$, vers l'espérance de chacune des variables. Au sens où pour chaque $\varepsilon > 0$, la probabilité de se tromper de plus de ε tend vers 0.

Evidemment, puisque toutes nos variables ont la même loi, on a aussi que $E[X_j^2] = E[X_1^2] < +\infty$ pour tout j. On a vu que si $E[X_1^2] < +\infty$, alors X_1 est d'espérance finie. Donc $E[X_1]$ a un sens, et comme précédemment $E[X_j] = E[X_1]$ pour tout j.

Pour estimer le membre de gauche de (6.19), on va appliquer l'inégalité de Bienaymé-Chebyshev à la variabble S_n. Par linéarité de l'espérance,

$$E[S_n] = \sum_{j=1}^{n} E[X_j] = nE[X_1].$$

Par le théorème 6.6 sur la variance d'une somme de variables indépendantes,

$$Var(S_n) = \sum_{j=1}^{n} Var(X_j) = nVar(X_1)$$

parce que les X_j on aussi la même variance. Et maintenant, par Bienaymé-Chebyshev

(6.20)

$$P\Big(\Big|\frac{S_n}{n} - E[X_1]\Big| \geq \varepsilon\Big) = P\Big(\Big|S_n - nE[X_1]\Big| \geq n\varepsilon\Big) = P\Big(\Big|S_n - E[S_n]\Big| > n\varepsilon\Big)$$
$$\leq \frac{Var(S_n)}{n^2\varepsilon^2} = \frac{Var(X_1)}{n\varepsilon^2}.$$

Donc non seulement le membre de gauche tend vers 0 quand n tend vers $+\infty$, mais on a une borne effective, qui naturellement dépend de ε, mais qui est plus utile que l'énoncé du théorème. □

Exemple. Supposons qu'on lance 10 dés parfaits. Quelle est la probabilité p d'obtenir une somme de 50 ou plus ? Comme l'espérance de S_n est 35, et que par symétrie l'espérance de faire 10 ou moins est la même, on trouve

$$p = \frac{1}{2}P\big(\big|S_n - E[S_n]\big| \geq 15\big) \leq \frac{1}{2}\frac{Var(X_1)}{10 \times 1,5^2},$$

où 1.5 est l'écart entre S_n/n et son espérance. La variance de X_1 est

$$Var(X_1) = \frac{1}{6}\sum_{l=1}^{6}(l - \frac{7}{2})^2 = \frac{1}{3}\big[(1/2)^2 + (3/2)^2 + (5/2)^2\big] = \frac{1}{12}(1+9+25) = \frac{35}{12}.$$

Donc fnalement (et sauf erreurs que vous me signalerez)

$$p = \frac{1}{2}\frac{1}{10}\frac{4}{9}\frac{35}{12} = \frac{7}{9 \times 12} \simeq 0,065.$$

Avec 100 lancers, et toujours une moyenne de 5 ou plus, on aurait eu la probabilité $p/10 \simeq 0,0065$. Ca n'est pas si faible qu'on aurait (peut-être) cru. Mais on voit quand même que c'est plus utilisable que la loi faible des grands nombres.

Evidemment, le fait qu'on appelle ceci loi faible des grands nombres laisse entende qu'il y a des résultats (en fait, plein de résultats) plus précis.

On voit aussi que le fait que les variables avaient toutes la même loi était juste une facilité de calcul. Par contre l'indépendance est très importante.

7 Statistiques

On a déjà vu un exemple qui se rapproche de ce qu'on veut faire. Je donne trois exemples de base, on peut en imaginer plein d'autres.

Exemple 1 (sondage). On a une population de N personnes qui s'aprêtent à voter, ou qui ont une caractéristique qu'on va chercher à évaluer en moyenne. Pour avoir un exemple un peu différent des suivants, au lieu d'un vote par oui ou par non, où le paramètre serait la proportion p de "oui", disons qu'on cherche des informations sur la taille des personnes, évaluée en centimètres (on va prendre des nombres entiers).

On se propose de fait un sondage sur n personnes, ce qui signifie qu'on va choisir n personnes parmi les N, et noter consciencieusement la taille de chacun des individus choisis. Ensuite on pourra en déduire les valeurs présumées de diverses quantités pour la population générale.

On voudrait par exemple évaluer la taille moyenne dans la population, ou éventuellement d'autres quantités comme l'écart type ou juste la proportion de la population qui fait plus de $170cm$. Disons qu'on choisit de deviner la taille moyenne. On voudrait aussi un intervalle de confiance, c'est-à-dire un intervalle I de valeurs de la quantité choisie tel qu'on soit sûr, disons à 5%, que la véritable valeur (la taille moyenne) est dans cet intervalle.

Exemple 2 (pièce lestée ?) On a une pièce, peut-être lestée, et on voudrait savoir si il est légitime de soupçonner qu'elle n'a pas "exactement" une chance sur deux de tomber sur pile. On lance la pièce n fois, et on voudrait savoir quelle conclusion on peut tirer des n lancers.

Exemple 3 (casino tricheur ?) Un casino prétend qu'avec la nouvelle ma-

chine qu'il vient de mettre en place les joueurs on une chance sur 3 de gagner. Une association de défense des consommateurs prétend que c'est faux, et évalue plutôt les chances de gagner à une sur 10. Est-ce qu'en jouant n fois, le juge a une grand probabilité de les mettre d'accord avec peu de chances d'erreur judiciaire ?

Exemple 4 (trop de cas de cancer ?) Un médecin trouve que dans sa région, il y a 1% de la population qui est malade d'un certain cancer, alors que dans la population générale c'est 10 fois moins. Est-ce imputable au hasard ? Je veux dire, quelle est la probabilité que ce soit imputable au hasard ?

Je prétends qu'avec ce qui précéde, on a essentiellement les moyens de répondre aux questions plus haut. La méthode pour faire est souvent la même, donc j'essaie de décrire comment on fait en principe.

7.1 On trouve un modèle

Ca peut être délicat, parce que c'est ici que vont se cacher toutes sortes d'hypothèses implicites. Par exemple, pour le casino, on sait déjà qu'on est seulement intéressé par les chances de gains, donc si on projette de jouer n fois, on utilera $\Omega = \{0,1\}^n$, muni de la probabilité produit de n mesures de probabilité de Bernouilli de paramètre $p \in [0,1]$. On ferait pareil pour les lancers de pièce, ou pour le médecin.

Déjà dans ces cas, on voit bien qu'on a décidé de s'intéresser à n variables aléatoires $X_1, \dots X_n$, qui dans notre cas sont à valeurs dans $\{0,1\}$, et dont on suppose qu'elles doivent être indépendantes. Donc en fait ce qu'on a, c'est la loi de chacune d'entre elles (ici, une loi de Bernoulli de paramètre p), comme on décide que ces variables sont indépendantes (parce que l'expérience est faite pour ça), et donc leur loi jointe est la loi produit, sur $\{0,1\}^n$. Une fois qu'on a décidé ceci, le choix de Ω le plus simple est de prendre carrément $\Omega = \{0,1\}^n$, muni de la probabilité produit (donc de la loi jointe des n variables). Bien sûr on aurait pu prendre Ω artificielllement plus grand, ou donner un autre nom aux éléments de Ω, mais on n'aurait sans doute fait que compliquer arbitrairement les choses. Dit d'une autre manière, ce qui nous intéresse dans Ω, c'est surtout la loi des variables X_j, qui de toute façon doit contenir toutes les informations utiles.

J'insiste qu'ici p (l'espérance de nos variables X_j) est un paramètre, qu'on ne connait pas encore, et qu'on va chercher à évaluer.

Passons au sondage pour voir si on a bien compris. Cette fois ce qu'on va retenir c'est la taille de chacun des membres de notre échantillon de n personne. Appelons X_j la taille de la j-ème personne. On va décider que c'est un nombre entier compris entre 1 et 250. On prévoit large, mais ça n'est pas grave. Cette fois, il y a plusieurs paramètre dans la description de notre modèle, qui sont la proportion de chacune des tailles possible dans notre population totale. Autrement dit, les données sont les nombres $p(k)$, où $1 \leq k \leq 250$ et $p(k)$ est la proportion de la population générale qui a la taille k. Dit autrement, si X_j est la taille de la j-ème personne de notre échantillon, on suppose que (par choix au hasard) toutes les X_j ont la même loi, et cette loi est donnée par $P(X_j = k) = p(k)$. Et du coup le plus raisonnable est de prendre $\Omega = E_{250}^n$, muni de la probabilité produit de n fois la loi de X_1 (donnée par les $p(k)$).

Dans tous ces cas, on s'est débrouillé pour que lles X_j soient les coordonnées dans Ω, ce qui simplifie les notations ! Dans notre cas des sondages, la loi est juste un peu plus compliquée, et est donnée par 249 paramètres (la somme est 1).

On aurait pu pareillement faire un sondage sur les trois personnalités préférées (parmi un groupe de 100), et pour chaque $j \leq n$ on aurait choisi un triplet de nombres inférieurs à 100. Donc par exemple, chaque $X_j(\omega)$ serait choisi dans E_{100}^3 (et en plus avec trois coordonnées différentes). Ensuite l'inconnue est la loi de chaque X_j (elles sont toujours de même loi et indépendantes), et l'espace Ω serait $[E_{100}^3]^n$, muni de la probabilité produit (n fois) de la loi inconnue de chaque X_j. Si on veut faire le malin, on peut décider de noter Z l'ensemble des triplets de E_{100}^3 dont les trois coordonnées sont différentes, et prendre $\Omega = Z^100$, toujours muni de la probabilité produit. Notre choix initial de Ω était plus grand, parce qu'on a ajouté des événements ω dont on sait bien qu'ils ne se produiront pas. Et bien sûr dans ce dernier cas, essayer d'évaluer la moyenne des X_j n'a aucun sens. Ca serait pareil pour un vote de premier tour.

En tout cas, à ce stade insistons qu'on a déjà décidé que notre expérience suit une loi bien précise, mais on ne connaît pas le paramètre p ou les divers paramètres, qui déterminent la loi des X_j, on essaie juste de deviner p, ou d'autres quantités dépendant des paramètres de la loi, avec une expérience

(comme interroger $n$personnes, ou jouer n fois, ou lancer n fois la pièce, on consulter les dossiers de ses n patients).

7.2 Estimateurs

On suppose donc maintenant qu'on s'est donné un modèle (donc un Ω et une mesure de probabilité P sur Ω) qui dépend d'un paramètre qu'on va appeler $\theta \in \Theta$. Le plus souvent dans nos exemple, ce paramètre est $p \in [0,1]$, mais comme dans le cas du sondage sur les tailles, Θ peut être plus compliqué, comme par exemple donné par la liste de tous les nombres $p(k)$, $1 \leq k \leq 250$.

On pense très fort que dans les cas ci-dessus, Ω est juste l'ensemble des valeurs que peut prendre le n-uplet des variables aléatoires X_j, que P est la loi jointe des X_j, et que Θ est juste un ensemble de paramètres qui décrivent la loi des X_j.

On essaie d'évaluer un des paramètres θ qui entrent dans la définition de Θ. Désolé, c'est compliqué parce que j'essaie de faire plusieurs cas à la fois. Dans les cas simples où P est juste donné par un paramètre p, le plus simple est de dire qu'on essaie d'évaluer p. Mais par exemple dans le cas du sondage, θ peut être la taille moyenne dans la population globale, ou le pourcentage de gens dont la taille est ≥ 159, ou la variance de la taille. On va se contenter du cas où $\theta \in \mathbb{R}$.

Définition 7.1. *Un estimateur $\widehat{\theta}$ de θ est une variable aléatoire $\omega \to \widehat{\theta}(\omega)$, définie sur Θ et à valeurs dans $\mathbb{R}$.*

Mince ! c'est une définition bizarre, et en plus elle ne dépend pas de θ. C'est comme ça ; par contre la notion d'un bon estimateur de θ dépendra de la définition de θ. Pour l'instant, je veux juste dire que $\widehat{\theta}$ est quelquechose que je calcule, d'une manière décidée à l'avance, à partir d'une formule qui fait intervenir ω. Dans le cas standard Ω est juste le produit des images des X_j, je veux dire que $\widehat{\theta}$ est juste calculé à partir des valeurs des X_j, et surtout pas de p ou θ qu'on ne connait pas. Bref, $\widehat{\theta}$ est quelquechose qu'on peut calculer à partir des résultats de l'expérience.

C'est la convention de noter les estimateurs de quantités comme θ avec un chapeau. Et bien que la définition ne le fasse pas apparaître clairement, en

pratique on choisit $\widehat{\theta}$ en fonction de la définition paramètre θ qu'on veut estimer et on dit que $\widehat{\theta}$ est un estimateur de θ. On dit aussi que la valeur $\widehat{\theta}(\omega)$ est une estimation de θ. Ce sera clair dans les exemples.

Prenons l'exemple le plus standard, celui de **l'estimateur de moyenne empirique.** On suppose qu'on a choisi nos variables aléatoires $X_1, \ldots X_n$ justement indépendantes et telles que $E[X_j] = \theta$. Je veux vraiment dire, telles que pour toutes valeurs que les paramètres à chercher puisse avoir, quand on fait l'expérience la moyenne (sur les valeurs possibles de $\omega \in \Omega$) de chacune de nos v.a. X_j est justement θ. C'est souvent ce qu'on organise. Par exemple, dans le cas du casino, on prend $\theta = p$ (la probabilité de gagner), et on choisit nos variables aléatoires X_j comme ci-dessus ($X_j = 0$ si on a perdu au coup j, 1 si on a gagné). Et par construction $E[X_j] = p$. Dans ce cas, la première chose qui vient à l'esprit est de choisir

$$\widehat{p} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n). \tag{7.1}$$

C'est ceci l'estimateur de moyenne empirique de $p = E[X_j]$.

Evidemment, il y a aussi de très mauvais estimateurs de θ. Par exemple prendre $\widehat{\theta} = 0$ quoiqu'il arrive est autorisé, c'est bien un estimateur de tout ce que vous voudrez, mais on ne recommande pas son usage parce sa valeur ne donnera aucune information sur l'expérience.

Fin du cours 11, 2020

Dans le cas du sondage sur la taille de la population, si vous voulez estimer la moyenne de taille (appelons-la m), l'estimateur empirique est bien $\frac{1}{n}(X_1+X_2+\ldots+X_n)$, puisque l'espérance de chaque X_j est bien m. Si vous voulez calculer le pourcentage q de la population qui mesure $170cm$, et que vous voulez vraiment l'écrire comme moyenne empirique, poser $Y_j(\omega) = 0$ quand $X_j(\omega) \neq 170$ et $Y_j(\omega) = 1$ quand $X_j(\omega) = 170$. Alors les Y_j sont indépendants (par exemple parce que les X_j le sont), et on peut vérifier que $E[Y_j] = q$. Du coup l'estimateur de q par la moyenne empirique des Y_j est $\widehat{q} = \frac{1}{n}(Y_1 + Y_2 + \ldots + Y_n)$.

Il y a des cas où il n'y a pas forcément une moyenne empirique qui saute aux yeux. Par exemple, si vous cherchez à estimer la moyenne M du carré de la taille, il n'est pas si facile a priori de définir des variables aléatoires X_j dont l'espérance est M, surtout que je vous rappelle que vous ne savez pas exactement qui est

la loi des X_j. Ca ne vous empêche pas d'essayer $\widehat{M} = \frac{1}{n}(X_1^2 + X_2^2 + \ldots + X_n^2)$, qui est raisonnable mais n'est pas une moyenne empirique.

Donc on peut dire qu'en général il y a une jungle d'estimateurs, et d'ailleurs certains seront meilleurs que d'autres pour certaines valeurs des paramètres Θ, alors que le contraire sera vrai pour d'autres valeurs. Certaines propriétés des estimateurs choisis sont rassurantes. En voici deux.

Définition 7.2. *Disons qu'on cherche à évaluer $\theta \in \mathbb{R}$. Un estimateur $\widehat{\theta}$ du paramètre θ est dit sans biais quand $E[\widehat{\theta}] = \theta$.*

Comprendre, pour tout choix de θ dans ce modèle, la formule qu'on a choisi donne un estimateur $\widehat{\theta}(\omega)$ qui a la bonne moyenne (quand on moyennise sur ω).

Et l'estimateur par la moyenne empirique ci-dessus est sans biais : $E[\widehat{\theta}] = \frac{1}{n}E[X_1 + \ldots + X_n] = \frac{1}{n}(E[X_1] + \ldots + E[X_n]) = \frac{n}{n}\theta = \theta$ par linéarité de l'espérance puis la définition. Il n'y a même pas besoin que les variables soient indépendantes.

Définition 7.3. *On cherche à évaluer $\theta \in \mathbb{R}$, à l'aide de n variables aléatoires X_j indépendantes et de même loi. Pour chaque n on se donne une formule $\widehat{\theta} = \widehat{\theta}_n(X_1, \ldots X_n)$. On dit que l'estimateur $\widehat{\theta}_n$ choisi ainsi est consistant quand pour tout $\varepsilon > 0$*

$$\lim_{n \to +\infty} P(|\widehat{\theta}_n - \theta| > \varepsilon) = 0. \tag{7.2}$$

Notez l'écriture assez implicite, comme d'habitude :

$$\big\{|\widehat{\theta}_n - \theta| > \varepsilon\big\} = \big\{\omega \in \Omega\,;\, |\widehat{\theta}_n(X_1(\omega), \ldots, X_n(\omega)) - \theta| > \varepsilon\big\}$$

est un événement, dont on calcule la probabilité.

Exemple typique : si les variables aléatoires X_j sont de carré sommable (on savait déjà espérance finie, et on ajoute variance finie), alors la moyenne empirique est un estimateur cohérent. C'est la loi faible des grand nombres. Et en plus, la démonstration avec Bienaymé-Chebyshev donne une idée de la vitesse à laquelle $P(|\widehat{\theta}_n - \theta| > \varepsilon)$ tend vers 0 : voir (6.20).

Remarque. Pour l'instant on n'a vu que des exemples où la moyenne empirique fait bien l'affaire, alors en voici un où un autre estimateur fait mieux le job. Je

suis spectateur du marathon de Paris (pour copier encore sur mon prédécesseur OH), et je voudrais savoir le nombre N d'inscrits. Je prends n de ces coureurs au hasard (on suppose que c'est possible, que le numéros ne trahissent pas grand-chose sur le coureur, et que les coureurs sont bien numérotés de 1 à N. Un moyen est de prendre la moyenne M des numéros repérés, de penser que la moyenne de tous les numéros doit être $(N+1)/2$, et de prendre $\widehat{N} = 2M - 1$ (de sorte que $(\widehat{N}+1)/2 = M$). Exercice : vérifier que c'est un estimateur sans biais. On vérifierait aussi, en regardant la démonstration de la loi des grands nombres, que la probabilité de se tromper de plus de 1%, par exemple, tend vers 0 quand n tend vers $+\infty$. Mais en fait, relativement lentement.

L'estimateur semblant plus débile qui consiste à retenir le plus grand des numéros de brassards observé, a en fait tendance à faire mieux. Pour que le sup des valeurs observées soit en-dessous de 99% de la valeur de N, il faut avoir eu un peu de malchance (en gros, une chance sur 100 à chaque coup), mais n fois de suite. Donc pour avoir raté son coup, c'est-à-dire atterri n fois de suite dans les 99%, est de l'ordre de $(\frac{99}{100})^n$, qui devient assez vite petit quand n est grand (plusieurs centaines). Cet estimateur n'est pas sans biais, et l'un des exos de TD montre comment on peut l'améliorer encore un peu. Mais il a aussi l'avantage de donner une borne inférieure complètement sure.

A ce stade on a dit comment essayer de deviner le paramètre θ (par exemple p dans les exemples de base), mais ce n'est pas satisfaisant parce qu'on n'a pas encore évalué les incertitudes et marges d'erreur. D'où les paragraphes suivants ou l'on continue la description du travail systématique d'un statisticien débutant !

7.3 Test d'hypothèse ; zone de rejet, niveau du test

Prenons l'exemple des n lancers de pièce. On fait une hypothèse qu'on va appeler H_0 : par exemple, p (la vraie probabilité) est égale à 1. On veut dire si c'est vraissemblable au vu de l'expérience qui suit.

On a déjà construit son modèle sur la loi des variables aléatoires X_j (les résultats de l'expérience qu'on va faire), puis on décide d'une zone de rejet de l'hypothèse. C'est un ensemble de valeurs pour les X_j pour lesquelles on décidera qu'on pense que H_0 est trop peu probable.

Par exemple, pour le cas des lancers de pièce, on se donne un intervalle $I = [a, b]$, et on décidera que la **zone de rejet** (on va l'appeler R_n) est quand $Z_n = (X_1 + \ldots + X_n)/n$ n'est pas dans $[a, b]$. On aurait pu imaginer une formule plus compliquée en fonction des X_j, mais utiliser les valeurs de la moyenne empirique (en fait Z_n est l'estimateur de moyenne empirique) est une bonne idée. Le plus raisonnable ici est de prendre un intervalle comme $I = [1-c, 1+c]$, avec c assez petit.

Ca c'est notre algorithme : on a choisi n et c, et on décide de rejeter H_0 quand (après avoir fait l'expérience) on trouve $\frac{1}{n}Z_n(\omega) \notin I$. On dit que le test qu'on vient de décrire est de **niveau** $\alpha \in [0, 1]$ quand, pour tout choix du vrai p, la probabilité de dénoncer H_0 à tort est au plus α. Dit autrement, puisqu'on ne peut pas rejeter à tort quand H_0 est fausse, si pour tout choix des paramètres pour lesquels θ vérifie H_0, on a

$$P(R_n) \leq \alpha. \tag{7.3}$$

J'ai dit choix des paramétres parce que dans le fond la loi des X_j peut dépendre d'autres paramètres que θ. Ca n'arrivera pas, et en tout cas dans le cas de la pièce lestée ou pas, le paramètre est $p \in [0, 1]$, et on a choisi de chercher $\theta = p$. Dans notre exemple précis, puisqu'on a dit que qu'on teste si $p = 1/2$ et qu'on rejette quand $S_n/n \notin I$, on voit que le test est de niveau α si

$$P(Z_n \notin I) \leq \alpha \text{ quand } p = 1/2. \tag{7.4}$$

On a déjà estimé cette probabilité en utilisant Bienaymé-Chebyshev comme dans la démonstration de la loi des grands nombres. En fait, si n n'est pas trop grand, on peut même faire (ou faire faire par un ordi) un calcul plus précis. On sait que nZ_n suit une loi binomiale $Bin(n, p)$; ici $p = 1/2$; on peut calculer $P(Z_n = \frac{k}{n}) = C_n^k p^k (1-p)^{n-k}$, avec $p = 1/2$, pour tout k, puis calculer exactement en fonction de c quelle est la probabilité que $Z_n \notin I$, et on en déduit le niveau du test très précisément. Quand n est trop gros pour qu'on puisse calculer, ou si on a une situation plus compliquée que des pièces, on se rabat sur BT comme ci-dessus.

Remarque de bon sens : prendre c plus grand nous permet de prendre α plus petit. Si on connait le niveau cherché α (suivant ce qu'on considère acceptable), et n, on pourra choisit c en conséquence (la plus petite valeur qui donne le bon

α à la fin). Et aussi, si on cherche un intervalle de confiance $I = [a, b]$ comme ci-dessus, plusieurs choix sont possibles (centré, pas centré, etc). D'autant plus que d'autres contraintes vont apparaître !

Les définitions générales sont comme le laisse entendre l'exemple. On a une loi sur Ω qui dépend de $\theta \in \Theta$. On rappellle qu'on ne connaît pas θ ; c'est justement lui qu'on essaie d'estimer.

On formule une hypothèse H_0. Ca veut dire qu'on se donne une partie Θ_0 de Θ, et que notre hypothèse est "$\theta \in \Theta_0$". Par exemple, on aurait pu choisir l'hypothése "$p \geq 1/3$" dans l'exemple du casino, qu'on peut exprimer comme "le casino ne triche pas". Ensuite on choisit n et une zone de rejet, c'est-à-dire un ensemble $R_n \subset \Omega$. Donc quand on fera l'expérience, on dira "je pense fort que le casino triche" si $\omega \in R_n$. Et enfin, on calcule le niveau du test choisi, qui est (l'infimum des) α tel que, comme en (7.4),

$$P(\omega \in R_n) \leq \alpha \text{ pour tout choix de } \theta \in \Theta_0. \tag{6}$$

On est obligé d'être un peu prudent ici, est d'envisager le cas (la valeur de θ) le pire pour notre test. Si les enjeux sont grands en cas de rejet injustifié (envoyer le pauvre gérant du casino en prison), on essaiera de choisir R_n pour que α soit assez petit ! C'est ce que j'appellerai le contrôle des erreurs de type 1.

Dans l'exemple du casino (et où $H_0 = \{p \geq 1/3\}$, il est raisonnable de prendre n assez grand, puis de définir R_n par $Z_n \geq \frac{1}{3} - c$ pour un $c > 0$, puis de calculer avec une loi binomiale la valeur de α obtenue. Noter que normalement il faudrait faire un calcul pour chaque $p \geq 1/3$ (pour chaque $\theta \in \Theta_0$ dans le cas général), mais heureusement on se convainc assez vite que la valeur de p qui donne la plus grande probabilité que H_0 soit rejetée à tort est $p = 1/3$. Donc en fait un seul calcul suffit dans ce cas.

Evidemment, plus on choisit c petit, plus il sera difficile de discriminer (et plus α aura des chances d'être trop grand. Dans le cas du casino, puisque l'association de consommateurs prétend que p est plutôt proche de $1/10$, prendre $c = 1/6$ et condamner le casino si on trouve $\alpha = 10^{-10}$ devrait satisfaire tout le monde (sauf le juge si ça l'oblige à jouer trop de fois ; faire le calcul avant de proposer l'expérience !).

Encore un commentaire : si l'association de consommateurs a calculé (sur la base d'expériences ci-dessus, pas par une indiscrétion sur le logiciel qui fait

marcher les machines à sou) que la probabilité de gagner n'était que de $0,332$, ne proposez pas l'expérience ci dessus : vous n'arriverez pas à distinguer de $1/3$. On va préciser un peu ceci dans le prochain paragraphe.

7.4 Contre-hypothèse H_1 et puissance du test

On continue sur l'histoire du casino : on se donne une second hypothèse H_1, disjointe de la première. En général, H_1 est une partie de Θ_0 qui donc ne rencontre pas Θ_0. Par exemple, dans le cas du casino, H_1 peut être "en fait $p \leq 1/10$" (comme le prétend l'association). On s'intéresse à la **puissance du test**, qui est

(7.5)
$$Puissance = \inf\left\{P(H_0 \text{ est rejetée}) \,;\, \theta \in \Theta_1\right\} = \inf\left\{P(\omega \in R_n) \,;\, \theta \in \Theta_1\right\}.$$

On veut une puissance la plus grande possible (proche de 1). Il s'agit de repérer les erreurs de type 2, où H_1 était vrai (et donc H_0 largement fausse), et où le test ne nous a pas donné d'information concluante.

A nouveau c'est normal de prendre la borne inférieure sur Θ_1, puisqu'on ne sait pas qui est θ, et qu'on veut des assurances sur le test quelque soit $\theta \in \Theta_1$.

Plus on prend Θ_1 loin de Θ_0, plus on a des chances de trouver un test puissant. Dans le cas du casino, si la probabilité de gagner est bien $1/10$ comme le disent les consommateurs, ca devrait être facile de démontrer que le casino triche ! Donc si le test que vous (le juge) avez choisi est puissant (disons 99% ou plus), et que vous ne rejetez pas H_0 finalement, les consommateurs auront du mal à expliquer que c'est une horrible erreur judiciaire ; ils n'avaient qu'à ne pas annoncer des résultats aussi outrageux).

A nouveau, dans le cas de nos exemples, les probabilités à calculer en (7.5) sont calculables à partir de tables de probabilités pour les lois binomiales. Et dans le cas du casino, il est facile de se convaincre que la probabilité la plus petite sera obtenue dans le cas où $p = 1/10$, ce qui permet d'estimer la puissance assez facilement. Je crois que (calculs recopiés d'OH) jouer une vingtaine de fois suffit déjà à avoir une très bonne idée de la situation !

Résumé de la méthode souvent employée. **Marche à suivre** pour des cas un peu plus généraux que ceux des exemples.

1. N'oubliez pas de **définir un modèle**, ce qui en pratique signifie un espace Ω (typiquement pour n expériences ; il faudra choisir n à un moment), muni de sa mesure de probabilité P. Le plus souvent, (Ω, P) sera donné par la loi jointe de variables aléatoires $X_1, \ldots, X_n$, une par expérience élémentaire.

Et d'ailleurs le plus souvent les variables seront indépendantes et de même loi, et donc la loi sera une loi produit.

2. Choisir une **hypothèse H_0 à tester**. Normalement H_0 est de la forme $\omega \in \Theta_0$. Par exemple, la probabilité p associée à la pièce est $1/2$, ou est dans un intervalle I_0 donné. Ou l'affirmation "pour au moins 80% de la polulation, la taille des indivdus est entre 160 et 185 cm." Cette affirmation dépend uniquement de la répartition des tailles dans la population totale, qui est notre liste de paramètres Θ.

3. Choisir, le cas échéant, une **contre-hypothèse (hypothèse alternative)** H_1. Elle vous servira à démontrer que votre test est bon, au sens où il discrimine bien entre Θ_0 et Θ_1. On vérifiera ceci en calculant la puissance.

4. Choisir une **statistique de test**. Dans nos cas standard, la statistique de test sera une fonction de nos variables $X_1, \ldots X_n$. Comme un estimateur de θ qui dépend de ces n valeurs, puisqu'après tout c'est θ qui nous intéresse. Par exemple, si on a l'occasion de le faire, prendre la moyenne des valeurs des X_j ; mais ca pourrait être autre chose. On dira "la moyenne empirique" quand en fait le nombre θ à estimer est l'espérance de chaque X_j. Ca arrivera souvent, parce qu'on essaiera souvent de choisir des X_j dont l'espérance est θ. Mais ça n'est pas une obligation.

5. Choisir une **zone de rejet**. Donc un ensemble R_n, en général de valeurs de ω, ou le plus souvent un sous-ensemble (notons le encore R_n de valeurs de la statistique choisie), où dans notre protocole on décidera que l'hypothèse H_0 est rejetée. Dans les cas simples comme celui du casino, ca sera un ou plusieurs intervalles de valeurs de $\widehat{p}$, et on rejettera si la valeur estimée est dans l'un de ces intervalles.

6. Calculer le **niveau α du test**, en fonction de ce qu'on a choisi avant (et de n). En fait c'est mieux de garder des paramètres dans la description des choix plus haut, parce qu'en fait on dit plutôt choisir le niveau α à l'avance et se débrouiller pour que nos choix (n en particulier) donnent cette (petite) valeur

de α.

Ce sont les contraintes du problème qui vous font choisir α : dire qu'on prévoit que le mauvais candidat sera élu, c'est moins grave que d'envoyer un innocent aux travaux forcés.

7. Le cas échéant, calculer **la puissance du test** en fonction du H_1 que vous vous êtes donné. C'est le meilleur moyen de vérifier si votre test est assez discriminant. Dans l'absolu, les erreurs des deux types doivent être petites : rejeter H_0 alors qu'il était vrai (condamner un innocent) et ne pas voir que H_0 est faux dans des cas où ca devrait être clair (innocenter un gros coupable). Ca peut être l'occasion de vous rendre compte avant de commencer que vous avez choisi n trop petit, ou une mauvaise zone de rejet.

Dans le cas du casino, vous pouvez éventuellement prendre comme hypothèse H_0 que $p = 1/3$ (on les prend au mot), et choisir la zone de rejet $\widehat{p} \in [1/2, 1]$. Ca sera un test avec petit α (tout va bien jusqu'ici), mais pas bien puissant : si comme le disent les consommateurs, la probabilité de gagner est très faible, votre estimateur ne tombera presque jamais dans la zone de rejet, donc vos chances de détecter une arnaque sont presque nulles. Votre test ne serait puissant qu'avec des machines programmées pour faire perdre le casino.

8. Ne pas se gourrer sur l'interprétation des résultats. Par exemple, si vous rejetez H_0 avec un petit α, vous pouvez dire : compte tenu du modèle que j'ai choisi pertinent parce que ...., je peux affirmer qu'il y a moins de 100α chances sur 100 pour que H_0 soit vrai. Donc comme c'est le gérant qui est responsable de la machine, le risque qu'il soit innocent est bien faible en regard du préjudice des sommes probablement détournées, et hop je l'envoie en prison.

Si vous ne rejetez pas, au maximum vous pouvez dire : je n'ai pas trouvé de preuve statistique que ce gérant de casino est coupable. Pourtant j'avais un test hyperpuissant, alors si les affirmations de l'association de consommateurs (avec les 10%) étaient exactes, j'aurais presque certainement découvert l'arnaque ; circulez, il n'y a rien à voir. Ou alors, plus modeste : il faut bien dire que mon test n'était pas bien puissant, mais je n'avais pas envie de passer les 10 prochains mois à jouer au bandit manchot. Alors tant pis j'acquitte.

7.5 Intervalles de confiance

Disons qu'on lance une pièce n fois et qu'on veut calculer sa probabilité p de faire face. Mais les mêmes calculs s'appliqueraient pour un sondage, ou des tests de dépistage aléatoire dans une population, ou plein d'autres choses semblables.

On a un estimateur raisonnable $\widehat{p} = \widehat{p}(\omega)$, à savoir la moyenne des n résultats $X_j(\omega)$ des n lancers. Qu'on a appelé moyenne empirique (on la calcule après les expériences).

Mais on en veut plus : est-ce qu'on a des chances de s'être beaucoup trompé. Donc, ce qui est mieux est de se donner une petite valeur de $\alpha \in]0,1[$ à l'avance, et de demander un ensemble $\widehat{I} = \widehat{I}(X_1, \dots X_n)$ (ici, le mieux sera de prendre un intervalle, et donc on veut qu'il soit calculé à partir des X_i) tel que

$$P(p \in \widehat{I}) \geq 1 - \alpha \quad \text{pour tout } p \in [0,1]. \tag{7.6}$$

Rappelons qu'ici $\widehat{I}$ dépend des valeurs $X_1, \dots X_n$ observées, et que la probabilité à calculer dépend aussi de p. C'est donc une affaire un peu compliquée.

Comme on ne sait pas à l'avance qui est p, il est logique de demander (7.6) pour tout p, parce que comme cela, pour le vrai p en particulier, on sait que la probabilité de tomber hors de l'intervalle de confiance était bien inférieure à α. Donc si on s'est trompé on n'a pas eu de chance. Ca ressemble un peu à la formule de Bayes, mais c'est différent : il n'y a ici qu'une seule vérité (p), mais on a α chances d'avoir été malheureux dans le tirage. Si on recommence, on aura toujours la même petite chance de se tromper.

Maintenant comment construire un intervalle de confiance et calculer le α qui correspond ? On se place en fait dans le cas où le paramètre cherché est p, on a tiré n variables de Bernoulli indépendantes X_j de paramètre p, et on veut utiliser les valeurs des X_j pour calculer un intervalle de confiance.

Il est raisonnable de prendre $I = [\widehat{p}-t, \widehat{p}+t]$, où l'on peut choisir la valeur de t en fonction de n et de la précision (du α) souhaitée. Comme on ne connait pas p, on va éviter de tout calculer avec des lois binomiales (on aurait trop de cas à regarder), et à la place on essaie d'estimer $|\widehat{p} - p|$ de manière aussi uniforme possible. Ca tombe bien, Bienaymé-Chebishev va nous aider. On note X_j nos variables de Bernoulli, et S leur somme. On sait que la somme a l'espérance np (puisque $E[X_j] = p$ pour tout j), et que sa variance est $np(1-p)$ (la somme

des variances des X_j). Rappelons que $\widehat{p} = S/n$ est la moyenne des X_j ; ainsi $E[\widehat{p}] = E[S]/n = p$ et $Var(\widehat{p}) = n^{-2}Var(S) = p(1-p)/n$. Maintenant

$$\begin{aligned} P(p \notin I) &= P(|\widehat{p} - p| \geq t) = P(|\widehat{p} - E[\widehat{p}]| \geq t) = P(|\widehat{p} - E[\widehat{p}]|^2 \geq t^2) \\ &\leq t^{-2}E[|\widehat{p} - E[\widehat{p}]|^2] = t^{-2}Var(\widehat{p}) = t^{-2}p(1-p)/n \leq \frac{1}{4t^2n} \end{aligned}$$

où l'on a utilisé Markov parce que c'est plus facile à retenir que BC, et sauf erreur de ma part. Heureusement que pour la dernière inégalité on a trouvé une borne uniforme indépendante de p.

Bon maintenant on veut choisir α, par exemple 10^{-2}. On veut aussi avoir une précision de 10^{-1} sur p, donc pouvoir prendre $t = 10^{-1}$. On dirait que ceci nous force à prendre n tel que $4 \cdot 10^{-2}n \geq 100$, donc $n \geq 2500$. Avec ce nombre de lancers, on aura la précision souhaité avec moins de 1% de chances de se tromper. C'est une peu fatigant, la loi des grands nombre ne donne pas une convergence rapide, mais c'est mieux que rien.

En termes de sondés, si on veut une précision de 1%, et avec une fiabilité de 99%, on dirait qu'il va falloir sonder 250000 personnes. Bien entendu notre estimation est améliorable (parce que BC est ici une inégalité stricte). Quand même, ca fait pas mal de personnes à sonder et ça vaut le coup de chercher des astuces (chercher la personne qui vote toujours comme la majorité, au risque de se tromper).

Noter que la question : combien de voix "oui" sur un sondage de 1000 personnes me faut-il pour pouvoir décider, avec fiabilité de 99%, que la réponse au référendum sera oui, est un peu plus simple et relève de la partie test avec "hypothèse H_0" faite plus haut. Et là cela vaut le coup d'utiliser les tables de la loi binomiale $Bin(1000, 1/2)$ pour optimiser le calcul.

Il est clair que ces informations ne feront pas de vous un statisticien aguerri, mais (la note optimiste ?) vous vous débrouillerez au moins nettement mieux que pas mal de gens entendus dans les médias récemment.