Université d'Avignon
et des Pays de Vaucluse

IUP Génie Informatique et Mathématique
Licence 2^{e}année

Introduction aux probabilités et aux statistiques

Denis Allard

Table des matières

I. PROBABILITÉS

Introduction

Considérons les énoncés suivants :
– « tirer un 6 sur un dé équilibré » ;
– « il pleut demain » ;
– « une météorite de grande taille va heurter la terre au cours du prochain siècle ».
Ce sont des énoncés de complexité croissante. Une probabilité est une façon d'assigner un nombre à de tels énoncés. En général, ce nombre est compris entre 0 et 1, mais parfois on l'exprime en pourcentage. Plus ce nombre est élevé, plus l'énoncé est considéré comme probable. Pour pouvoir associer un nombre à ces énoncés, il faut tout d'abord préciser ce qu'on entend par un énoncé, qu'on appelle *évènement* en langage probabiliste. Puis il faut se donner des règles sur la manière de calculer ce nombre. C'est l'objet de la théorie des probabilités présentée dans la permière partie de ce cours.
Il faut toujours avoir présent à l'esprit qu'une probabilité ne provient que d'une théorie mathématique et non directement de la réalité. La modélisation probabiliste sert à pallier : notre absence de connaissance (pour le loto, le poker, les cours de la bourse du lendemain, la charge sur un serveur,...), une trop grande complexité du phénomène étudié (en économétrie, évolution du climat,...), ou d'une connaissance imparfaite des conditions initiales d'un phénomène dont par ailleurs on connaît les équations (lancer de dé, météo,...). Il faut toutefois noter un cas très particulier où les probabilités semblent irréductibles et sont pour l'instant considérées comme inhérentes à la nature même de l'objet étudié : la mécanique quantique.

Les probabilités servent également de base théorique aux statistiques, dont l'objet est la collecte, la description et l'analyse de données.

Ce cours présente les notions fondamentales des probabilités, avec comme fil conducteur la nécessité d'arriver aux notions de probabilité nécessaires aux statistiques : variables aléatoires, loi des grands nombres et théorème central limite. On commence par poser les axiomes des probabilités et rappeler quelques notions indispensables de la théorie des ensembles (chapitre 1). On présente ensuite le modèle de probabilité uniforme que l'on rencontre dans les jeux de dés et de cartes (chapitre 2). Au chapitre 3 on aborde le cœur du raisonnement probabiliste avec les probabilités conditionnelles et la célèbre formule de Bayes. Les chapitres 4 à 6 sont consacrés aux variables aléatoires. Le chapitre 7 présente les notions liées au calcul de l'espérance : espérance mathématique, variance et covariance. Enfin, le chapitre 8 est consacré aux théorèmes limites qui sont indispensables à l'utilisation des probabilités en statistique.

1 Evènements, Ensembles et Probabilités

1.1 Les évènements

En probabilité, un évènement est un énoncé en relation avec l'expérience aléatoire considérée. Ainsi, parler de l'âge du capitaine quand on joue à la roulette ne peut pas être un évènement. Les définitions suivantes servent à construire la notion d'évènement.

Une **expérience aléatoire** est une expérience dont on ne connaît pas l'issue : lancer de dé, distribution d'un jeu de cartes, temps d'attente à un guichet de poste, temps de téléchargement d'un gros fichier,...

L' **univers** est l'ensemble de tous les résultats possibles d'une expérience aléatoire. L'univers est noté Ω.

Un **évènement élémentaire** est un des résultats possibles d'une expérience aléatoire. Un évènement élémentaire est aussi appelé un **état**, noté ω. On note donc

$$\omega \in \Omega.$$

Un **évènement** est un ensemble de résultats possibles, c'est-à-dire un ensemble d'états. Les évènements sont généralement notés par les premières lettres capitales romaines, $A, B, C, D, \ldots$

Un énoncé se traduit par un évènement, qui est un sous-ensemble de l'univers.

Quelques exemples :

- On joue deux fois à Pile ou Face. Un état est par exemple (P, F). L'univers est

  $$\Omega = \{(P,P);(P,F);(F,P);(F,F)\}.$$

  Il y a quatre états dans l'univers. A l'énoncé « tirer au moins une fois pile » correspond l'évènement $A = \{(P,P);(P,F);(F,P)\}$. Il y a trois états dans cet évènement.
- On lance deux dés. Un état est par exemple (3,6). L'univers est

  $$\Omega = \{(i,j) : i,j \in \{1,2,3,4,5,6\}\}$$

  Il y a 36 états dans cet univers. A l'énoncé « la somme des deux dés vaut 7 » correspond l'évènement $A = \{(1,6);(2,5);(3,4);(4,3);(5,2);(6,1)\}$. Il y 6 états dans cet évènement.
- Tirer à Pile ou Face jusqu'à ce que Face apparaisse. Un état est par exemple $PPPPF$. L'univers contient un nombre infini (mais dénombrable) d'éléments :

  $$\Omega = \{F, PF, PPF, PPPF, PPPPF, \ldots\}.$$

  A l'énoncé « Face apparaît en 3 lancers au plus », correspond l'évènement $A = \{F, PF, PPF\}$.

– On observe la durée de vie d'un transistor. Un état peut être n'importe quel réel positif, par exemple 3 h 48 mn et 5,666... s. Ici l'univers est $\mathbf{R}^+$. C'est un ensemble continu, c'est à dire non dénombrable. A l'énoncé « Le transistor fonctionne encore après 60 jours » correspond l'évènement $]7200\,\mathrm{h}, +\infty[$. Cet évènement n'est pas non plus dénombrable.

1.2 Rappels sur la théorie des ensembles

Nous avons vu que les univers sont des ensembles et les évènements des sous-ensembles (donc des ensembles eux-mêmes) de cet univers. Il est donc nécessaire de rappeler ici quelques notions sur la théorie des ensembles.

Définitions

– L'**intersection** de deux ensembles A et B est l'ensemble noté $A \cap B$ des éléments ω qui appartiennent à A **et** à B.
– L'**union** de deux ensembles A et B est l'ensemble noté $A \cup B$ des éléments ω qui appartiennent à A **ou** à B.
– Un ensemble A est **inclus** dans B, et on note $A \subset B$, si tout élément ω appartenant à A appartient aussi à B.
– Le **complémentaire** (sous-entendu par rapport à Ω) d'un ensemble A, est l'ensemble noté $\bar{A}$ (parfois noté A^c) des éléments de Ω n'appartenant pas à A.
– Les intersections et unions se généralisent au cas d'ensembles multiples. Ainsi par exemple,

$$\bigcup_{i=1}^{n} A_i = \{\omega : \omega \in A_1, \text{ ou } \omega \in A_2, \text{ ou } \omega \in A_3, \ldots, \text{ou } \omega \in A_n\}$$

désigne la réunion des ensembles $A_1, \ldots, A_n$. On écrit de même

$$\bigcap_{i=1}^{n} A_i = \{\omega : \omega \in A_1, \text{ et } \omega \in A_2, \text{ et } \omega \in A_3, \ldots, \text{et } \omega \in A_n\}$$

pour l'intersection des ensembles $A_1, \ldots, A_n$.
Pour des séquences infinies d'ensembles, on fait de même.
– Les ensembles A et B sont dits **disjoints** (ou **incompatibles**) si leur intersection est vide : $A \cap B = \emptyset$.

On rappelle les propriétés suivantes, qu'il est facile de vérifier, à l'aide des diagrammes de Venn par exemple.

– **Commutativité :**

$$\begin{aligned} A \cap B &= B \cap A \\ A \cup B &= B \cup A \end{aligned}$$

– **Associativité :**

$$\begin{aligned} A\cap(B\cap C) &= (A\cap B)\cap C &= A\cap B\cap C \\ A\cup(B\cup C) &= (A\cup B)\cup C &= A\cup B\cup C \end{aligned}$$

– **Distributivité :**

$$\begin{aligned} A\cap(B\cup C) &= (A\cap B)\cup(A\cap C) \\ A\cup(B\cap C) &= (A\cup B)\cap(A\cup C) \end{aligned}$$

– **Lois de De Morgan :**

$$\begin{aligned} \overline{A\cap B} &= \bar{A}\cup\bar{B} \\ \overline{A\cup B} &= \bar{A}\cap\bar{B} \end{aligned}$$

Ainsi, le contraire de « la chaîne hi-fi fonctionne » est « la platine ne fonctionne pas, **ou** l'ampli ne fonctionne pas, **ou** les enceintes ne fonctionnent pas ».
Le contraire de « la chaîne hi-fi ne fonctionne pas » est « la platine fonctionne, **et** l'ampli fonctionne, **et** les enceintes fonctionnent ».

1.3 La tribu des évènements

Au paragraphe précédent, nous avons vu qu'à des énoncés correspondent des sous-ensembles de Ω, les évènements. Composer des énoncés à l'aide des conjonctions « et », « ou » et à l'aide de la négation revient donc à composer des ensembles à l'aide des opérateurs $\cap$, $\cup$ et $\bar{}$. Lorsque l'on compose des énoncés (en français) valables, on souhaite qu'ils restent des énoncés valables. De même, lorsque l'on compose des évènements, on souhaite qu'ils restent des évènements. On notera $\mathcal{A}$ l'ensemble des évènements que l'on souhaite prendre en compte dans le cadre d'une expérience aléatoire. La structure algébrique qui correspond à ce que nous souhaitons s'appelle une **tribu**. Elle possède les propriétés suivantes, dont on pourra s'assurer qu'elles correspondent bien à ce que nous recherchons.

Définition 1.1 *La* **tribu** $\mathcal{A}$ *des évènements (sous-entendu, d'une certaine expérience aléatoire dont l'univers est* Ω*) est définie par les deux propriétés suivantes :*
1) Si $A\in\mathcal{A}$*, alors* $\bar{A}\in\mathcal{A}$ *;*
2) Si $A,B\in\mathcal{A}$*, alors* $A\cup B\in\mathcal{A}$*.*

Cette structure suffit si Ω possède un nombre fini d'états. Lorsque ce n'est pas le cas, on remplace la propriété 2) par la propriété 2') :

2') Pour une famille (éventuellement infinie dénombrable) d'évènements $(A_n)_{n\geq 0}$ *de* $\mathcal{A}$*, on a*

$$\bigcup_{n=0}^{\infty} A_n \in \mathcal{A}.$$

Il faut noter que l'axiome 2) ou 2') entraîne nécessairement que $\Omega \in \mathcal{A}$. C'est l'évènement certain. En conséquence, l'axiome 1) entraîne que $\emptyset \in \mathcal{A}$ également : c'est l'évènement impossible. Le couple $(\Omega, \mathcal{A})$ s'appelle un **espace probabilisable**. En effet, à ce stade, nous avons construit un espace en attente d'être probabilisé en précisant sur quel ensemble d'évènements il est licite de calculer des probabilités. En revanche, nous n'avons encore rien dit sur la *manière* de calculer ces probabilités, et nous verrons au paragraphe suivant qu'il y en a de multiples.

Voici deux cas particuliers très importants :

1. Si Ω est fini, alors $\mathcal{A} = \mathcal{P}(\Omega)$, l'ensemble des parties de Ω, est une tribu. C'est la tribu la plus détaillée (ou dit aussi la plus fine) que l'on puisse construire sur Ω. C'est un peu la « tribu par défaut ».
2. Si Ω est un intervalle I de $\mathbf{R}$ (par exemple $I = \mathbf{R}^+$), alors la tribu engendrée par les intervalles ouverts inclus dans I est la tribu « naturelle ». On l'appelle la tribu borélienne[1].

1.4 Probabilités

Ayant un espace probabilisable $(\Omega, \mathcal{A})$, comment associer un nombre à chaque évènement ? En d'autres termes, comment construire une application qui à chaque évènement A associe un nombre $P(A)$?

Avant de répondre directement, faisons un peu d'histoire des probabilités. Jusqu'au XIX$^{\text{e}}$siècle, on définissait la probabilité de A comme la limite de la fréquence de A quand l'expérience aléatoire était répétée un très grand nombre de fois : $P(A) = \lim_{n\to\infty} N(A)/n$. En effet, quand on reporte $N(A)/n$ en fonction de n, on observe une courbe qui oscille de moins en moins pour se stabiliser autour d'une certaine valeur. Cette définition posait de très grosses difficultés car :

– Observer la convergence peut être long. Tant que l'on converge vers des valeurs simples facilement calculables, tout va bien. C'était souvent le cas, car aux XVII$^{\text{e}}$et XVIII$^{\text{e}}$siècles (Pascal, Leibniz, Moivre, Bernoulli) la notion de probabilité s'est essentiellement dégagée à partir de calculs sur les jeux de cartes et de dés pour lesquels l'univers n'admettait qu'un nombre fini d'états et présentait des caractères de symétrie. Mais quand on ne connaît pas la valeur limite, comment être certain de la valeur vers laquelle la courbe semble converger ?

[1] Une « tribu engendrée » se construit de la façon suivante : tous les ouverts $]a, b[$ de I sont des éléments de la tribu ; puis on construit les autres éléments de la tribu par les opérations de complémentation et réunions dénombrables sur des éléments de la tribu.

– Elle ne donne aucune règle pour faire des calculs sur des évènements complexes. Il fallait faire appel au bon sens (qui est souvent trompeur) et il s'ensuivait des débats acharnés entre probabilistes.

Vers 1910, le mathématicien russe Kolmogorov proposa une base axiomatique des probabilités, détachée de toute notion fréquentiste. C'est toujours cette définition qui est utilisée aujourd'hui. Elle a révolutionné les probabilités en offrant un cadre rigoureux dans lequel les calculs se font sans ambiguités. Il partit du principe qu'une probabilité assigne un nombre à un évènement : c'est donc une application. Il établit ensuite que l'ensemble des évènements doit avoir la structure algébrique d'une tribu. Enfin il donna les conditions que doit vérifier cette application pour être une probabilité. On les appelle les axiomes (de Kolmogorov) des probabilités.

Définition 1.2 (Axiomes de Kolmogorov) *Une probabilité P est une application de $\mathcal{A}$ vers $[0,1]$ vérifiant les 3 axiomes suivants :*

1. *la probabilité de l'évènement certain est 1 :*
$$P(\Omega) = 1;$$
2. *les probabilités de deux évènements disjoints s'ajoutent :*
$$\text{si } A \cap B = \emptyset, \text{ alors } P(A \cup B) = P(A) + P(B);$$
3. *Si Ω est infini, on complète l'axiome 2 par l'axiome 3 : si $(A_n)_{n\geq 0}$ est une famille d'évènements deux à deux incompatibles (c'est-à-dire, $A_n \cap A_m = \emptyset,\ \forall n, m$), alors*
$$P(\bigcup_{n=0}^{+\infty} A_n) = \sum_{n=0}^{+\infty} P(A_n).$$

Le triplet $(\Omega, \mathcal{A}, P)$ est un espace probabilisé. Beaucoup de probabilités P différentes peuvent correspondre à un espace probabilisable : p.ex. pour une pièce de monnaie non équilibrée, on peut prendre n'importe quelle valeur $0 \leq p \leq 1$ pour $p = P(Face)$. La forme particulière que prend l'application P est ce qu'on appelle un *modèle probabiliste* et les probabilités sont une théorie mathématique permettant de manipuler ces applications P.

Il faut noter que les fréquences statistiques vérifient bien les axiomes des probabilités. En effet :

1. $N(A,n)/n \in [0,1]$;
2. $N(\Omega,n)/n = n/n = 1$;
3. $N(\cup_i A_i, n)/n = \sum_i N(A_i,n)/n$ si les A_i sont disjoints.

Nous verrons au chapitre 8 que les 3 axiomes des probabilités (et quelques notions introduites plus tard) suffisent à démontrer le bien-fondé de l'approche fréquentiste, c'est-à-dire que $N(A)/n$ converge bien vers une valeur limite, égale à $P(A)$.

Voyons maintenant quelques propriétés qui découlent (presque) directement des 3 axiomes. Les démonstrations les plus simples sont laissées à titre d'exercice.

1. $P(\bar{A}) = 1 - P(A)$ (utiliser que A et $\bar{A}$ sont disjoints puis l'axiome 2) ;
2. Si $A \subset B$, alors $P(A) \leq P(B)$ (construire des évènements disjoints) ;
3. $0 \leq P(A) \leq 1$ (utiliser les deux premières propriétés).
4. **Formule de Poincaré :** Soient A et B deux évènements quelconques. On a :

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Preuve : Comme $A \cup B = A \cup (B \cap \bar{A})$ et que A et $B \cap \bar{A}$ sont disjoints, alors

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

Comme d'autre part $B = (B \cap A) \cup (B \cap \bar{A})$, on a

$$P(B) = P(B \cap A) + P(B \cap \bar{A}).$$

En rapprochant les deux équations, on trouve le résultat recherché. □.

Cette formule peut se généraliser à n évènements, en se compliquant fortement tout de même.

5. Soit une famille infinie d'évènements$(A_n)_{n>0}$ avec $A_1 \subset A_2 \subset A_3 \subset \ldots$, et $A = \cup_{n=1}^{\infty} A_n$. Alors $P(A_n) \to P(A)$ lorsque $n \to \infty$.

Preuve : On note $B_n = A_n \cup \bar{A}_{n-1} = A_n \backslash A_{n-1}$ (avec $A_0 = \emptyset$). Les B_n sont disjoints, et $\cup_{n=1}^{\infty} B_n = A$. Donc,

$$P(A_m) = P(\cup_{n=1}^{m} B_n) = \sum_{n=1}^{m} P(B_m).$$

Lorsque $m \to \infty$,

$$\lim P(A_m) = \lim \sum_{n=1}^{m} P(B_m) = P(A).$$

□

2 Probabilité uniforme - Dénombrement

2.1 Définition de la probabilité uniforme

A un espace probabilisable donné, on peut associer plusieurs probabilités. Le choix de la probabilité se fait à partir de la connaissance qu'on a du phénomène que l'on cherche à modéliser. Dans ce chapitre, nous allons nous intéresser à une probabilité particulière, la probabilité uniforme, qui est celle où tous les états (ou évènements élémentaires) sont équiprobables. Cette probabilité apparaît pour des raisons de symétrie (dés, cartes, boules dans une urne,...) ou lorsqu'on n'a aucune bonne raison d'utiliser un autre modèle (c'est un peu le modèle « par défaut »).

Définition 2.1 *Soit $\Omega = \{\omega_1, \omega_2, \ldots, \omega_{N_\Omega}\}$ un univers* **fini** *contenant N_Ω éléments. La probabilité P est la probabilité uniforme si tous les états ω_i, $i = 1, \ldots, N_\Omega$ sont équiprobables. Alors :*

$$P(\{\omega_1\}) = P(\{\omega_2\}) = \cdots = P(\{\omega_{N_\Omega}\}) = p,$$

avec $p = 1/N_\Omega$.

En effet, $1 = P(\Omega) = P(\cup_i \{\omega_i\}) = \sum_i P(\{\omega_i\}) = pN_\Omega$ par l'axiome 2. Dans ce modèle, la probabilité de tout évènement A se calcule facilement. Soit $A = \{\omega_1, \ldots, \omega_{N_A}\}$ un évènement contenant N_A éléments. Alors,

$$P(A) = \sum_{\omega_i \in A} P(\omega_i) = p.N_A = \frac{N_A}{N_\Omega}.$$

Habituellement, on appelle N_Ω le nombre de cas possibles et N_A le nombre de cas favorables (par rapport à l'évènement A), ce qui mène à la formule bien connue

$$P(A) = \frac{\text{nombre de cas favorables}}{\text{nombre de cas possibles}}.$$

Exemple On joue 3 fois à Pile ou Face. $A =$« Face apparaît une fois exactement ». Alors :

$$A = \{(FPP), (PFP), (PPF)\},$$

donc $N_A = 3$, et

$$\Omega = \{(FFF), (FFP), (FPF), (FPP), (PFF), (PFP), (PPF), (PPP)\},$$

et $N_\Omega = 8$. Donc $P(A) = 3/8$.

Ainsi, dans le cas d'une probabilité uniforme, le calcul des probabilités se ramène à compter le nombre des cas favorables et possibles. C'est ce qu'on appelle le dénombrement.

2.2 Règles de dénombrement

Dénombrer des cas favorables et des cas possibles n'est pas toujours simple. Ce paragraphe présente quelques règles de dénombrement qui seront nécessaires pour la suite. Nous commençons par des exemples introductifs, qui sont généralisés ensuite par des règles ou des théorèmes.

Exemple 1 Un restaurant propose une formule dans laquelle on peut choisir une entrée, un plat et un dessert. Il y a un total de 3 entrées, 5 plats et 4 desserts. Combien de repas différents peut-on prendre ? Pour chaque choix d'entrée, il y 5 plats, ce qui nous donne $3 \times 5 = 15$ combinaisons entrée+plat. Pour chacune des ces combinaisons, il y a 4 desserts, ce qui nous fait $15 \times 4 = 60$ menus complets différents. Au total, il y a donc $3 \times 5 \times 4$ menus.

Exemple 2 On joue 3 fois à Pile ou Face. Il y a $2 \times 2 \times 2 = 8$ résultats possibles à ce jeu (voir le paragraphe précédent pour plus de détail).

Ces deux exemples se généralisent par la règle de multiplication :

Règle de multiplication : lorsqu'une expérience est composée de m sous-expériences, et si chacune possède n_k résultats possibles quels que soient les résultats des autres sous-expériences, alors le nombre total de résultats possibles est :

$$n = n_1 n_2 \dots n_m.$$

On peut multiplier les exemples d'application de cette règle : plaque d'immatriculation, codes CB, mots de passe,...

Exemple 3 Soit $E = \{a, b, c\}$. Les permutations sont au nombre de 6 :

$$abc,\ acb, bac,\ bca,\ cab,\ cba.$$

Permutations : Le nombre de permutations d'un ensemble à n éléménts est $n!$ (rappel : « factorielle n » est le nombre $n! = n(n-1)\dots 1$.). C'est toujours la règle de multiplication qui s'applique (trouver pourquoi). Il faut noter que $n!$ croît très vite :

$1! = 1$	$5! = 120$	$9! = 362\,880$
$2! = 2$	$6! = 720$	$10! = 3\,628\,800$
$3! = 6$	$7! = 5\,040$	$70! > 10^{100}$
$4! = 24$	$8! = 40\,320$	

Par convention $0! = 1$.

Exemple 4 Douze personnes appartiennent à une association. Combien de bureaux composés d'un président, d'un vice-président, d'un trésorier et d'un secrétaire peut-on constituer ? On applique la règle de multiplication, ce qui nous donne 12 présidents $\times$ 11 vice-présidents $\times$ 10 trésoriers $\times$ 9 secrétaires. La réponse est donc 12.11.10.9. Or :

$$12.11.10.9 = 12.11.10.9.\frac{8!}{8!} = \frac{12!}{(12-4)!}.$$

Ce résultat se généralise en donnant la notion d'arrangement.

Arrangements : Lorsqu'on extrait k éléments d'un ensemble de n éléments, et que l'ordre dans lequel sont extraits ces éléments *importe*, on parle d'arrangements. On note A_n^k le nombre d'arrangements de k éléments parmi n ($k \leq n$), et

$$A_n^k = \frac{n!}{(n-k)!}.$$

Exemple 5 Supposons maintenant que dans cette association, le bureau est toujours composé de 4 personnes, mais aux fonctions équivalentes et interchangeables. Supposons que $a, b, c, d, \ldots, k, l$ soient les noms des douze membres de l'association. Alors (a, c, e, j), (a, e, c, j), (c, e, a, j), etc... sont des arrangements équivalents pour cette association, ainsi que toutes les 4! permutations de ce bureau. Si on souhaite compter le nombre de bureaux, il nous faut diviser le nombre A_n^k par le nombre de permutations de k éléments. C'est ce qu'on appelle les combinaisons.

Combinaisons : Lorsqu'on extrait k éléments d'un ensemble de n éléments, et que l'ordre dans lequel sont extraits ces éléments *n'importe pas*, on parle de combinaisons. On note C_n^k le nombre de combinaisons de k éléments parmi n ($k \leq n$), et

$$C_n^k = \frac{A_n^k}{k!} = \frac{n!}{k!(n-k)!}.$$

Quelques propriétés :

1. $$C_n^{n-k} = \frac{n!}{(n-k)!(n-(n-k))!} = \frac{n!}{(n-k)!k!} = C_n^k,$$
   ce qui revient à exprimer l'évidence suivante : sélectionner k éléments parmi n est équivalent à sélectionner les $n-k$ complémentaires.
2. Il n'y a qu'un seule manière de sélectionner tous les éléments d'un ensemble :
   $$C_n^n = \frac{n!}{n!(n-n)!} = \frac{n!}{n!0!} = 1.$$
   En combinant avec le résultat précédent : $C_n^0 = 1$.
3. Il y a n façon de sélectionner un élément parmi n :
   $$C_n^1 = \frac{n!}{1!(n-1)!} = n.$$

Les exemples d'application des combinaisons sont multiples. Ainsi par exemple :
- On peut choisir 4 toppings parmi 15 pour une pizza :
  cela nous donne $C_{15}^4 = 15.14.13.12/(4.3.2) = 1365$ pizzas différentes.
- Au poker on donne 5 cartes parmi 32 :
  cela nous donne $C_{32}^5 = 32.31.30.29.28/(5.4.3.2) = 201\,376$ mains différentes.
- Au loto, il faut sélectionner 6 numéros parmi 49 :
  ce qui nous donne $49.48.47.46.45.44/(6.5.4.3.2) = 13\,983\,816$ grilles différentes.

2.3 Le binôme de Newton

Les combinaisons C_n^k sont parfois appelées *coefficients du binôme*, car ils interviennent dans le développement du binôme

$$(x+y)^n.$$

Théorème 2.1 (Formule du binôme) *Pour tous réels x et y et pour tout entier naturel n, on a :*

$$(x+y)^n = \sum_{k=0}^{n} C_n^k x^k y^{n-k}.$$

Nous donnerons deux preuves à ce théorème. La première est basée sur un argument combinatoire, en lien avec le paragraphe précédent. La seconde démonstration, plus classique, est purement algébrique et sert à convaincre ceux qui « ne croient pas » à la première.

Preuve 1 : C'est un polynôme de degré n, homogène en x et y. Donc après développement en somme de monômes, chaque monôme doit être de degré n, donc de la forme $\alpha_{k,n} x^k y^{n-k}$. Il reste à déterminer la valeur de $\alpha_{k,n}$. Ce terme s'obtient en choisissant k facteurs pour x (et donc $n-k$ facteurs pour y). Nous venons de voir qu'il y a C_n^k façons de choisir k facteurs parmi n. Donc $\alpha_{k,n} = C_n^k$. □.

Preuve 2 : On démontre par récurrence. Il faut donc i) démontrer que la formule du binôme est vraie pour $n = 1$, puis ii) démontrer qu'elle est vraie pour $n+1$ quand on l'admet pour n.

i)

$$\begin{aligned}(x+y)^1 &= x+y \\ &= C_1^0 x^1 y^0 + C_1^1 x^0 y^1.\end{aligned}$$

ii)

$$\begin{aligned}(x+y)^{n+1} &= (x+y)(x+y)^n \\ &= (x+y)\sum_{k=0}^{n} C_n^k x^k y^{n-k} \\ &= \sum_{k=0}^{n} C_n^k x^{k+1} y^{n-k} + \sum_{k=0}^{n} C_n^k x^k y^{n+1-k}.\end{aligned}$$

On fait le changement de variable $i = k+1$ dans la première somme et $i = k$ dans la seconde, ce qui nous donne

$$\begin{aligned}&= \sum_{i=1}^{n+1} C_n^{i-1} x^i y^{n+1-i} + \sum_{i=0}^{n} C_n^i x^i y^{n+1-i} \\ &= x^{n+1} + \sum_{i=1}^{n} x^i y^{n+1-i}\left(C_n^{i-1} + C_n^i\right) + y^{n+1}. \qquad (1)\end{aligned}$$

Or :

$$\begin{aligned} C_n^{i-1} + C_n^i &= \frac{n!}{(i-1)!(n-i+1)!} + \frac{n!}{i!(n-i)!} \\ &= \frac{i.n! + (n-i+1)!n!}{i!(n+1-i)!} \\ &= \frac{n!(i+n-i+1)}{i!(n+1-i)!} = \frac{(n+1)!}{i!(n+1-i)!} \\ &= C_{n+1}^i. \end{aligned}$$

En remplaçant ce résultat dans (1), on trouve :

$$\begin{aligned} (x+y)^{n+1} &= x^{n+1} + \sum_{i=1}^{n} x^i y^{n+1-i} C_{n+1}^i + y^{n+1} \\ &= \sum_{i=0}^{n+1} x^i y^{n+1-i} C_{n+1}^i, \end{aligned}$$

ce qui démontre la proposition. □

Notons qu'au passage on a également démontré que :

$$C_n^{k-1} + C_n^k = C_{n+1}^k.$$

Cette égalité peut se représenter graphiquement par le **triangle de Pascal**[2], dans lequel on retrouve à la ligne n et à la colonne k la valeur C_n^k :

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
$n=1$	1	1				
$n=2$	1	2	1			
$n=3$	1	3	3	1		
$n=4$	1	4	$\boxed{6}$	$\boxed{4}$	1	
$n=5$	1	5	10	$\boxed{10}$	5	1

[2]Newton (1642–1727) et Pascal (1623–1662) : ce ne sont pas là des mathématiques récentes !

3 Probabilités conditionnelles – Indépendance

3.1 Probabilité conditionnelle

Supposons maintenant que pour évaluer la probabilité d'un évènement A, on sache qu'un autre évènement, B, s'est produit. Il est possible que connaître cette information augmente la probabilité de A, ou au contraire la diminue, ou encore la laisse inchangée. On appelle probabilité conditionnelle de A sachant B cette nouvelle probabilité, notée $P(A \mid B)$. Avant de présenter la définition de cette probabilité conditionnelle, voyons un exemple.

Exemple 1 On lance deux dés équilibrés. On considère $A =$ « somme des dés vaut 8 » et $B =$ « le premier dé est un 3 ». Dans ce cas, $A = \{(2,6);(3,5);(4,4);(5,3);(6,2)\}$, donc $P(A) = 5/36$, et $B = \{(3,1);(3,2);(3,3);(3,4);(3,5);(3,6)\}$. On cherche la probabilité conditionnelle $P(A \mid B)$. Si A se produit sachant que B se produise, il faut nécessairement que les deux se produisent, et donc le seul état « favorable »qui corresponde à cela est $(3,5)$. Mais, si nous savons que B s'est produit, le nombre d'états « possibles » n'est plus 36, mais 6, et dès lors la probabilité est $P(A \mid B) = 1/6$.

Cela nous amène à la définition suivante :

Définition 3.1 *Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé et soient A et B deux évènements aléatoires avec $P(B) \neq 0$. On appelle* **probabilité conditionnelle** *de A sachant B, le rapport*

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

Comme dans l'exemple introductif, cette définition s'explique de la façon suivante : au numérateur, comme A et B se produisent simultanément, c'est bien $P(A \cap B)$ qu'il faut calculer ; au dénominateur, $P(B)$ joue le rôle de constante de normalisation, ou, si l'on préfère, comme on sait que B s'est produit, il joue le rôle d'univers restreint, c'est à dire de nouvel évènement certain. Ainsi, on s'assure que $P(B \mid B) = 1$.

Exemple 2 Mme Durand a 2 enfants, dont l'un au moins est un garçon. Quelle est la probabilité qu'elle ait deux garçons (on suppose les évènements élémentaires équiprobables) ?
Notons $A =$ « avoir deux garçons » et $B =$ « avoir au moins un garçon ». On recherche $P(A \mid B)$. Alors $\Omega = \{FG, FF, GF, GG\}$, $A = \{GG\}$, $B = \{FG, GF, GG\}$ et $A \cap B = \{GG\}$. On trouve les probabilités suivantes : $P(B) = 3/4$, $P(A \cap B) = 1/4$ et $P(A \mid B) = P(A \cap B)/P(B) = 0.25/0.75 = 1/3$. Une réponse naïve (et fausse) aurait consisté à dire que si un des enfants est un garçon, il y a une chance sur deux pour que l'autre soit aussi un garçon. Poser les calculs à l'aide d'évènements bien définis permet d'éviter cette erreur.

Théorème 3.1 *Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé. Soient A et B deux évènements aléatoires avec $P(B) \neq 0$. Considérons l'application $A \mapsto P(A \mid B)$, et notons la $P_B(A)$. Alors $P_B(A)$ est bien une probabilité, au sens des axiomes de Kolmogorov.*

Preuve : La preuve consiste à démontrer que les axiomes de Kolmogorov (Définition 1.2) sont bien vérifiés pour P_B.

1. $P_B(A) \geq 0$, car $P(A) \geq 0$.
2. $$P_B(\Omega) = \frac{P(A \cap \Omega)}{P(A)} = \frac{P(A)}{P(A)} = 1.$$
3. Pour une suite A_i d'évènements disjoints,

$$\begin{aligned} P_B(\cup_i A_i) &= P(\cup_i A_i \mid B) = \frac{P((\cup_i A_i) \cap B)}{P(B)} \\ &= \sum_i \frac{P(A_i \cap B)}{P(B)} = \sum_i P(A_i \mid B) \\ &= \sum_i P_B(A_i). \end{aligned}$$

□

On en tire immédiatement les corollaires suivants :
– $P(\bar{A} \mid B) = 1 - P(A \mid B)$.
– $P(\emptyset \mid B) = 0$.
– $P(A_1 \cup A_2 \mid B) = P(A_1 \mid B) + P(A_2 \mid B) - P(A_1 \cap A_2 \mid B)$.

On peut faire une lecture inversée de la définition d'une probabilité conditionnelle, qui nous donne la **formule des probabilités composées**

$$P(A \cap B) = P(A \mid B)P(B).$$

En fait, on utilise très souvent les probabilités conditionnelles pour calculer des probabilités d'intersection en utilisant cette formule.

Exemple 3 Quelle est la probabilité de tirer deux as d'un jeu de 52 cartes ? Notons
$B_1 =$ « première carte est un as » et $B_2 =$ « deuxième carte est un as »
Alors, $A = B_1 \cap B_2$, et

$$P(A) = P(B_1 \cap B_2) = P(B_2 \mid B_1)P(B_1) = \frac{3}{51}\frac{4}{52}.$$

Notons que nous serions arrivés au même résultat par dénombrement : $P(A) = C_4^2/C_{52}^2 = 4.3/52.51$.

3.2 Indépendance

Définition 3.2 *Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé. Deux évènements sont* **indépendants** *si $P(A \cap B) = P(A)P(B)$.*

Exemple 1 On tire au hasard une carte d'un jeu de 52 cartes. On considère les évènements $A =$ « c'est un as » et $B =$ « c'est un pique ». Le modèle de probabilité est celui des états équiprobables. On trouve donc immédiatement que $P(A) = 4/52 = 1/13$ et $P(B) = 13/52 = 1/4$. D'autre part $P(A \cap B) = 1/52 = P(A)P(B)$. Donc les évènements A et B sont indépendants.

Exemple 2 On lance deux dés équilibrés. On considère les évènements $A =$ « la somme vaut neuf » et $B =$ « le premier dé est un 4 ». Ici également le modèle est celui des états équiprobables. Alors $P(A) = 4/36 = 1/9$ et $P(B) = 1/6$. D'autre part $P(A \cap B) = 1/36 \neq P(A)P(B)$. Donc les évènements ne sont pas indépendants. Considérons maintenant l'évènement $C =$ «le premier dé est un 2». Alors $P(C) = 1/6$. Mais, $P(A \cap C) = 0 \neq P(A)P(C)$. Les évènements A et C ne sont donc pas non plus indépendants.
Ce dernier résultat mène à une remarque très importante, qui généralise cet exemple :

Il ne faut pas confondre évènements indépendants et évènements incompatibles.

L'incompatibilité signifie que $A \cap B = \emptyset$ (il n'y a donc pas de notion de probabilité), tandis que l'indépendance signifie que $P(A \cap B) = P(A)P(B)$. En fait, deux évènements de probabilités non nulles ne peuvent être en même temps incompatibles et indépendants. La preuve est immédiate, car d'une part $P(A)P(B) > 0$ par hypothèse, mais d'autre part $A \cap B = \emptyset$, et donc $P(A \cap B) = P(\emptyset) = 0$. Ainsi, $P(A)P(B) \neq P(A \cap B)$. En particulier, A et $\bar{A}$ ne sont pas indépendants.

Théorème 3.2 *Soient A et B deux évènements d'un espace probabilisé $(\Omega, \mathcal{A}, P)$ avec $P(A) > 0$ et $P(B) > 0$. Alors, $P(A \mid B) = P(A) \Leftrightarrow (A, B)$ sont indépendants. De plus, dans ce cas, $(\bar{A}, B)$, $(A, \bar{B})$, $(\bar{A}, \bar{B})$ sont également indépendants.*

Preuve :
$\Rightarrow$) Par hypothèse, $P(A) = P(A \mid B) = P(A \cap B)/P(B)$. Donc $P(A)P(B) = P(A \cap B)$.
$\Leftarrow$) $P(A \mid B) = P(A \cap B)P(B) = P(A)P(B)/P(B) = P(A)$. □

L'interprétation que l'on peut faire de cette propriété est que si l'information apportée par B dans $P(A \mid B)$ ne change pas la probabilité $P(A)$, c'est que les deux évènements sont indépendants.

Définition 3.3 *Soit une famille d'évènements $A_1, A_2, \ldots, A_n$ d'un espace probabilisé $(\Omega, \mathcal{A}, P)$.*
- *Ces évènements sont dits indépendants deux à deux si pour toutes paires $\{A_i, A_j\}$ d'évènements, on a $P(A_i \cap A_j) = P(A_i)P(A_j)$.*
- *Ces évènements sont dits indépendants si pour tout sous-ensemble d'évènements on a $P(\cap_k A_k) = \prod_k P(A_k)$.*

On commet souvent l'erreur de penser que des évènements sont indépendants dès lors qu'ils sont indépendants deux à deux. En réalité, la seconde notion est beaucoup plus restrictive que la première.

Exemple 3 Soient les évènements $A =$ « Alice et Béatrice ont même anniversaire », $B =$ « Béatrice et Caroline ont même anniversaire » et $C =$ « Alice et Caroline ont même anniversaire ». On voit que $P(A) = P(B) = P(C) = 1/365$, et que $P(A \cap B) = P(B \cap C) = P(A \cap C) = 1/365^2$. Comme $P(A \cap B \cap C) = P(A \cap B) \neq P(A)P(B)P(C)$, les évènements A, B, C ne sont pas indépendants.

3.3 Probabilités totales

Considérons que les évènements $A_1, A_2, \ldots, A_n$ forment une partition de Ω, c'est-à-dire que leur réunion forme Ω, et que leur intersection deux à deux est vide :

$$\bigcup_{i=1}^{n} A_i = \Omega \quad \text{et} \quad A_i \cap A_j = \emptyset, \ \forall i \neq j.$$

Par exemple, si A est un évènement de Ω, $(A, \bar{A})$ est une partition de Ω. Alors,

$$\begin{aligned} B = B \cap \Omega &= B \cap (\cup_i A_i) \\ &= \cup_i (B \cap A_i). \end{aligned}$$

Donc,

$$\begin{aligned} P(B) &= P(\cup_i (B \cap A_i)) \\ &= \sum_i P(B \cap A_i) \\ P(B) &= \sum_i P(B \mid A_i) P(A_i). \end{aligned}$$

C'est la **formule des probabilités totales et composées**. L'avant-dernière égalité provient de l'axiome 2, en raison de l'incompatibilité des évènements formant la partition. La dernière égalité provient de la formule des probabilités composées. Cette formule permet de calculer une probabilité complexe en la décomposant sur une base d'évènements formant une partition, et menant à un ensemble de probabilités plus simples à calculer.

Exemple 1 Albert lance 2 pièces ; Béatrice lance 3 pièces. Gagne celui qui a le plus de Pile. On cherche à calculer $P(B)$, la probabilité que Béatrice gagne. On note les évènements suivants :
$A_i =$ « Albert a i Pile »,
$C_j =$ « Béatrice a j Pile ».

On décompose $P(B)$ de la façon suivante :

$$P(B) = \sum_{i=0}^{2} P(B \mid A_i) P(A_i).$$

Il faut donc évaluer les 6 probabilités de cette expression : On a $P(A_0) = 1/4$, $P(A_1) = 1/2$ et $P(A_2) = 1/4$, et
$P(B \mid A_0) = 1 - P(C_0) = 1 - 1/8 = 7/8$,
$P(B \mid A_1) = P(C_2 \cup C_3) = 1/8 + 3/8 = 4/8$,
$P(B \mid A_2) = P(C_3) = 1/8$,
ce qui donne au total $P(B) = 7/8.1/4 + 4/8.1/2 + 1/8.1/4 = (7 + 8 + 1)/32 = 1/2$.

Exemple 2 En génétique, chaque parent possède deux copies d'un gène, sur chacun des brins d'un chromosome. Les gamètes (les cellules de reproduction) ne possèdent chacune qu'une copie de chaque gène, qui provient d'un des deux brins, choisi au hasard avec une probabilité 1/2. Le génotype d'un individu est donc composé d'une copie du gène provenant de chaque parent, cette copie ayant été tirée au hasard parmi les deux copies de chacun des parents.

Supposons que le phénotype (ce que l'on *voit* de la personne, la couleur des yeux par exemple) soit déterminé par un seul gène, qui n'existe qu'en deux versions (les allèles) notées A et a. L'exemple historique est celui des petits pois étudiés par Mendel, où $A =$ « peau lisse » et $a =$ « peau ridée ». Chez les petits pois, la peau lisse est dominante, ce qui signifie que pour les combinaisons d'allèles AA, Aa et aA, le phénotype est une peau lisse. La peau ridée n'est obtenue que pour la combinaison aa. Chez les humains, la couleur bleue des yeux est dominée par la couleur marron. Soit une population dont les proportions sont les suivantes : $P(AA) = \alpha_0$, $P(Aa \text{ ou } aA) = \beta_0$ et $P(aa) = \gamma_0$. On suppose maintenant que les individus se croisent au hasard dans cette population, et on calcule les proportions des génotypes à la génération suivante. Soit $P(A)$ la probabilité que l'allèle A d'un ascendant soit sélectionné. En utilisant la formule des probabilités totales et composées, on a

$$\begin{aligned} P(A) &= P(A \mid AA)P(AA) + P(A \mid Aa \text{ ou } aA)P(aA) + P(A \mid aa)P(aa) \\ &= \alpha_0 + \beta_0/2 + 0 = p_1. \end{aligned}$$

Alors on obtient la fréquence de la nouvelle génération $P(AA) = P(A)P(A) = p_1^2 = \alpha_1$. De la même façon, on trouve $P(aA) = 2p_1(1-p_1) = \beta_1$ et $p(aa) = (1-p_1)^2 = \gamma_1$. Passons maintenant à la deuxième génération :

$$p_2 = \alpha_1 + \beta_1/2 = p_1^2 + 2p_1(1-p_1)/2 = p_1^2 + p_1 - p_1^2 = p_1.$$

Comme on vient de montrer que $p_2 = p_1$, il en découle immédiatement que $\alpha_2 = \alpha_1$, $\beta_2 = \beta_1$ et $\gamma_2 = \gamma_1$. Autrement dit, la proportion des différents génotypes (et donc aussi des phénotypes) reste identique dès la première génération, et cela quelle que soit la génération de départ. On peut donc écrire que $P(AA) = p^2$, $P(aA) = 2p(1-p)$ et $P(aa) = (1-p)^2$, où p est la proportion de l'allèle A dans la population. Ainsi, par exemple si Mendel observe 9% de petits pois avec la peau fripée, alors $P(aa) = 0,09 \Rightarrow 1 - p = 0,3$ et $P(A) = 0,7$.

3.4 Formule de Bayes

Supposons que l'on soit dans la situation suivante : on souhaite calculer une probabilité $P(B \mid A)$, mais on ne connaît que $P(A \mid B)$, $P(A \mid \bar{B})$ et $P(B)$. Cette situation se rencontre lorsque A est un effet que l'on observe, et qu'il y a deux causes possibles : B et $\bar{B}$. Il arrive souvent que l'on sache calculer la probabilité d'observer un effet pour une cause donnée (c'est-à-dire $P(A \mid B)$), et que l'on cherche à identifier la probabilité d'une cause, ayant observé l'effet.
On part de

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

Nous avons réussi à retourner le conditionnement, et on voit que $P(B \mid A)$ est proportionnel à $P(A \mid B)$. On sait calculer le numérateur. Il reste à calculer le dénominateur, ce qui se fait en utilisant la formule des probabilités totales et composées :

$$P(A) = P(A \mid B)P(B) + P(A \mid \bar{B})P(\bar{B}),$$

ce qui nous mène finalement à

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid \bar{B})P(\bar{B})}.$$

Cette formule se généralise à une partition $B_1, \ldots, B_n$ pour aboutir à la **formule de Bayes :**

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{n} P(A \mid B_j)P(B_j)}.$$

Exemple 1 Un constructeur d'ordinateurs fait venir ses CPU de trois usines différentes. L'usine B_1 construit 50% des CPU. Le taux de défaut est de $0,1\%$. L'usine B_2 fournit 30% des CPU dont $0,2\%$ défectueuses. La troisième usine B_3 fournit 20% des composantes avec un taux de défaut de $0,5\%$. On constate une CPU défectueuse sur un ordinateur. Quelle est la probabilité qu'elle vienne de la première usine, de la seconde, de la troisième ?

On note A l'évènement « puce défectueuse ». Le dénominateur vaut :

$$\begin{aligned} P(A) &= P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + P(A \mid B_3)P(B_3) \\ &= 0,1.0,5/100 + 0,2.0,3/100 + 0,5.0,2/100 = 0,21/100. \end{aligned}$$

D'où

$$\begin{aligned} P(B_1 \mid A) &= \frac{P(A|B_1)P(B_1)}{P(A)} = \frac{0,05/100}{0,21/100} = 5/21, \\ P(B_2 \mid A) &= \frac{P(A|B_2)P(B_2)}{P(A)} = \frac{0,06/100}{0,21/100} = 6/21, \\ P(B_3 \mid A) &= \frac{P(A|B_3)P(B_3)}{P(A)} = \frac{0,10/100}{0,21/100} = 10/21. \end{aligned}$$

On note que la somme de ces trois probabilités fait 1, puisque la CPU défectueuse doit bien provenir de l'une des trois usines.

Exemple 2 Un test sanguin détecte 98% des malades porteurs d'une maladie rare (prévalence dans la population = 1/1000). En revanche, 1% de la population saine est détectée positive (ce sont les faux positifs). Une personne est déclarée positive ; quelle est la probabilité qu'elle soit réellement atteinte ?

On note les évènements $A =$ « la personne est malade » et $B =$ « la personne est positive ». On veut calculer $P(A \mid B)$. On applique la formule de Bayes :

$$\begin{aligned} P(A \mid B) &= \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A})} \\ &= \frac{0,98 \times 0,001}{0,98 \times 0,001 + 0,01 \times 0,999} \\ &= 0,089. \end{aligned}$$

Dans cet exemple, une personne positive n'a donc qu'une probabilité de 9% d'être malade. Pour expliquer ce résultat qui paraît étonnant, considérons un ensemble de 1000 personnes tirées au hasard. Sur les 1000 personnes, 1 est malade et elle est presque toujours détectée. Sur les 999 personnes saines, 1%, c'est-à-dire environ 10 personnes seront déclarées positives. Au total, environ 11 personnes sont positives, dont une seule (soit $1/11 \simeq 9\%$) est malade.

Face à une probabilité aussi faible, lorsque une personne est déclarée positive, le test est refait pour confirmation. Des examens approfondis ne sont prescrits que si la personne est déclarée positive 2 fois. La probabilité qu'elle soit malade dans ce cas, $P(A \mid BB)$ vaut maintenant

$$P(A \mid BB) = \frac{P(BB \mid A)P(A)}{P(BB \mid A)P(A) + P(BB \mid \bar{A})P(\bar{A})}.$$

Afin de calculer les probabilités conditionnelles $P(BB \mid A)$ et $P(BB \mid \bar{A})$, on va supposer que les résultats de deux tests successifs sont indépendants[3]. Alors,

$$\begin{aligned} P(A \mid BB) &= \frac{P(BB \mid A)P(A)}{P(BB \mid A)P(A) + P(BB \mid \bar{A})P(\bar{A})} \\ &= \frac{0,98^2.0,001}{0,98^2.0,001 + 0,01^2.0,999} \\ &= 0,90. \end{aligned}$$

Ainsi, on voit que faire le test deux fois améliore très significativement les performances du test.

[3] Cette hypothèse sous-entend implicitement que les causes à l'origine du test erroné sont extérieures à l'individu et ne sont pas dues à des facteurs biologiques de l'individu, qui se répéteraient sur des tests successifs. Cette hypothèse est en réalité trop grossière pour être exacte en pratique.

4 Variables aléatoires

4.1 Introduction

Il arrive très souvent que la caractéristique qui nous intéresse à l'issue d'une expérience aléatoire soit un nombre. Ce nombre est variable puisqu'il varie d'une répétition à l'autre de l'expérience et il est aléatoire : on l'appelle donc **variable aléatoire**. Ainsi, par exemple :
– on tire à Pile ou Face 10 fois, on compte le nombre de Pile ;
– on lance deux dés, et on s'intéresse à la somme.

Définition 4.1 *Une* **variable aléatoire**, X, *est une application de* Ω *vers un ensemble quantitatif, le plus souvent* $\mathbf{N}$, $\mathbf{R}$ *ou une partie de ceux-ci,*

$$\begin{aligned} X \quad : \quad & \Omega \to E \\ & \omega \mapsto X(\omega) \end{aligned}$$

telle que pour tout intervalle I *de* E, *l'ensemble réciproque*

$$X^{-1}(I) = \{\omega \in \Omega \ : \ X(\omega) \in I\}$$

est un évènement de $\mathcal{A}$.

Afin d'alléger les notations, on parle alors de l'évènement $\{X \in I\}$, qui est une façon abrégée d'écrire l'évènement $\{\omega \ : \ X(\omega) \in I\}$. Le plus souvent, on ne fait plus référence à Ω, et on parle directement de $P(X \in I)$ plutôt que de $P(X^{-1}(I))$.

Il y a deux grands types de variables aléatoires :
– les variables aléatoires discrètes, qui ne prennent qu'un nombre fini ou infini dénombrable de valeurs ; le plus souvent $E = \mathbf{N}$, ou une partie de $\mathbf{N}$;
– Les variables aléatoires continues, qui prennent un nombre infini non dénombrable de valeurs ; le plus souvent $E = \mathbf{R}$ ou un intervalle de $\mathbf{R}$.

4.2 Fonction de répartition

Définition 4.2 *La fonction de répartition d'une variable aléatoire* X *(discrète ou continue) est la fonction* $F(x)$ *:*

$$F(x) = P(X \leq x), \quad x \in \mathbf{R}.$$

La fonction de répartition caractérise entièrement la variable aléatoire. Elle possède les propriétés suivantes :

i) $F(x)$ est non décroissante : si $x_1 < x_2$, alors $F(x_1) \leq F(x_2)$. Preuve :

$$\begin{aligned} F(x_2) &= P(X \leq x_2) \\ &= P(\{X \leq x_1\} \cup \{x_1 < X \leq x_2\}) \\ &= P(X \leq x_1) + P(x_1 < X \leq x_2) \\ &= F(x_1) + P(x_1 < X \leq x_2). \end{aligned}$$

Comme $P(x_1 < X \leq x_2) \geq 0$, $F(x_2) \geq F(x_1)$.

ii) $\lim_{x \to -\infty} F(x) = 0$.

iii) $\lim_{x \to \infty} F(x) = 1$.

iv) $F(x)$ est continue à droite : $\lim_{y \downarrow x} F(y) = F(x)$.

v) $\lim_{y \downarrow x} F(y) - \lim_{y \uparrow x} F(y) = P(X = x)$. Le saut de discontinuité de F au point x est égal à $P(X = x)$.

Remarque : En utilisant la fonction de répartition :

$$P(a < X \leq b) = F(b) - F(a).$$

En effet,

$$\begin{aligned} F(b) &= P(X \leq b) = P(\{X \leq a\} \cup \{a < X \leq b\}) \\ &= F(a) + P(a < X \leq b). \end{aligned}$$

4.3 Variables aléatoires discrètes

Une variable aléatoire discrète ne peut prendre qu'un nombre fini ou infini dénombrable de valeurs, qui seront notées $x_1, x_2, \ldots, x_n, \ldots$. L'ensemble de ces valeurs sera noté $\mathcal{X}$. On appelle **loi de probabilité**, ou **distribution** de X l'application p_X définie par

$$p_X(x_i) = P(X = x_i),$$

pour tout $x_i \in \mathcal{X}$. Pour abréger les notations, on écrit souvent $p_i = P(X = x_i)$. Une loi de probabilité est habituellement représentée par un diagramme en bâton.

Théorème 4.1 *Une famille discrète de valeurs p_i, $i = 1, 2, \ldots$, est une loi de probabilité si elle vérifie les deux conditions suivantes :*

$$\begin{aligned} p_i &\geq 0 \quad \forall i \\ \sum_i p_i &= 1. \end{aligned}$$

La preuve se fait en montrant que ces conditions vérifient les axiomes de Kolmogorov, et est immédiate. La seconde condition découle du deuxième axiome, en observant que

$$\Omega = \bigcup_i \{X = x_i\}.$$

Exemple 1 On lance deux dés. X est la somme des deux faces visibles et Y désigne le maximum de ces deux faces. Les lois de probabilité sont données dans le tableau ci-dessous :

k	1	2	3	4	5	6	7	8	9	10	11	12
$p_X(k)$	–	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36
$p_Y(k)$	1/36	3/36	5/36	7/36	9/36	11/36	–	–	–	–	–	–

Les propriétés iv) et v) des fonctions de répartition montrent que la fonction de répartition d'une variable aléatoire discrète est une fonction en escalier dont les sauts se font aux valeurs prises par la variable aléatoire, et dont l'amplitude du saut est la probabilité attachée à cette valeur.

Exemple 2 Considérons maintenant que X possède la distribution de probabilité suvante : $P(X = 0) = 1/8$, $P(X = 1) = 3/8$, $P(X = 2) = 3/8$ et $P(X = 3) = 1/8$. Calculons quelques valeurs particulières de la fonction de répartition :

$F(1,3) = P(X \leq 1,3) = P(X = 0) + P(X = 1) = 1/8 + 3/8 = 1/2$
$F(-0,3) = P(X \leq -0,3) = 0$
$F(0) = P(X \leq 0) = P(X = 0) = 1/8$
$F(0,00001) = P(X \leq 0,00001) = P(X = 0) = 1/8$
$F(0,99999) = P(X \leq 0,99999) = P(X = 0) = 1/8$
$F(1) = P(X \leq 1) = P(X = 0) + P(X = 1) = 1/8 + 3/8$
etc...

4.4 Modélisation probabiliste discrète

Dans ce paragraphe, nous allons voir quelques exemples de variables aléatoires discrètes, qui sont parmi les plus utilisées en modélisation probabiliste[4].

Variable aléatoire uniforme discrète

Une variable aléatoire discrète est dite uniforme si toutes les valeurs sont équiprobables :

$$P(X = x) = \frac{1}{n+1}, \quad k \leq x \leq k+n,$$

[4] Attention, dans ce paragraphe, et à plusieurs reprises dans ce cours, les termes « loi » et « variable aléatoire » sont souvent employés indifféremment. Le concept de loi de probabilité est plus opératoire, car c'est celui qui nous permet de faire les calculs, et son utilisation tend en conséquence à se généraliser. En toute rigueur, il s'agit d'un abus de langage, mais en général le contexte permet de comprendre de quoi on parle.

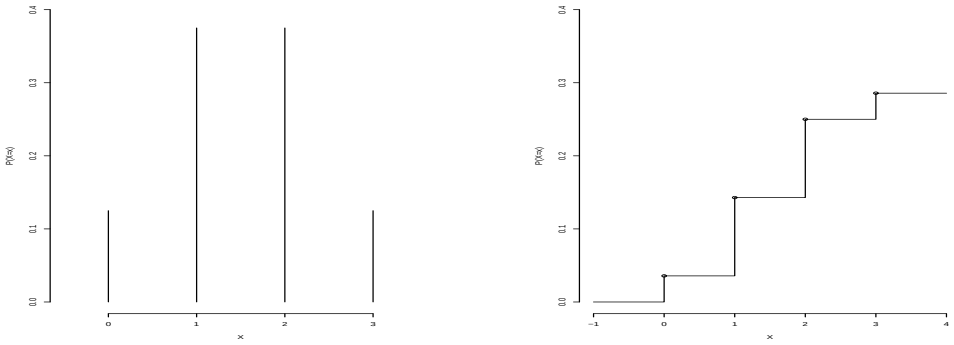

FIG. 1 – Système de probabilité (à gauche) et fonction de répartition (à droite) d'une variable aléatoire binomiale $\mathcal{B}(3, 1/2)$.

où k et $n \in \mathbf{N}$.

Exemple On lance un dé à six faces. X est la valeur de la face visible. Alors X est une variable aléatoire discrète entre 1 et 6, donc $k = 1$ et $n = 5$.

Variable aléatoire de Bernoulli

La variable aléatoire de Bernoulli est la variable aléatoire non triviale la plus simple possible. Elle ne prend que deux valeurs, le plus souvent 0 et 1. Elle sert très souvent de variable indicatrice pour ce qu'on appelle une **épreuve de Bernoulli**. Une épreuve de Bernoulli est une expérience aléatoire dont l'univers ne contient que deux états, le succès, noté A, et l'échec, noté $\bar{A}$. On peut associer une variable aléatoire X qui prend la valeur 1 lorsque l'épreuve est un succès, et la valeur 0 lorsque c'est un échec : c'est alors une variable aléatoire de Bernoulli. On note p la probabilité

$$p = P(X = 1) = P(A)$$

et $q = 1 - p$ la probabilité complémentaire. Nécessairement, on a $0 \leq p \leq 1$ (pourquoi ?).

Variable aléatoire géométrique

On répète une expérience aléatoire qui a une probabilité p de succès. On note X la variable aléatoire égale au nombre d'essais nécessaires pour observer le premier succès. Le système de probabilité de cette variable aléatoire est

$$P(X = k) = (1 - p)^{k-1} p, \qquad k = 1, 2, \ldots$$

En effet, pour que $X = k$, il est nécessaire d'observer $k - 1$ échecs successifs puis un succès au $k$ième essai. Par définition, X est une variable aléatoire géométrique de paramètre p (le système

de probabilité est une suite géométrique, d'où le nom de la variable aléatoire). Notons que ici $\mathcal{X} = \mathbf{N}^*$ est infini.
On peut vérifier que les conditions du théorème 4.1 sont bien vérifiées. En effet, $P(X = k) \geq 0$ et :

$$\begin{aligned}\sum_{k=1}^{\infty} P(X=k) &= p\sum_{k=1}^{\infty}(1-p)^{k-1} \\ &= p\sum_{k=0}^{\infty}(1-p)^{k} = p\frac{1}{1-(1-p)} = 1.\end{aligned}$$

Notons le résultat suivant :

$$\begin{aligned}P(X \geq k) &= \sum_{i=1}^{k} P(X=i) = p\sum_{i=1}^{k}(1-p)^{i-1} \\ &= p\,\frac{(1-p)^{k-1}}{p} = (1-p)^{k-1} \\ &= P(k-1 \text{ échecs successifs}).\end{aligned}$$

Variable aléatoire binomiale

On répète n fois une épreuve de Bernoulli ayant une probabilité de succès p. On désigne par X le nombre total de succès. Bien entendu, on a $0 \leq X \leq n$. La loi de probabilité est :

$$P(X = k) = p^k(1-p)^{n-k}C_n^k, \quad k = 0, \ldots, n.$$

Par définition, X est une variable aléatoire binomiale, de paramètres (n, p), ce qu'on note,

$$X \sim \mathcal{B}(n, p),$$

et qu'on lit : X est distribué selon une loi binomiale de paramètres (n, p). La forme de cette loi de probabilité paraît compliquée, mais elle s'explique assez facilement si on revient à son interprétation. Lorsqu'on répète n fois l'expérience aléatoire, une série de k succès et $n-k$ échecs est par exemple $AA\bar{A}A\bar{A}\bar{A}\ldots A\bar{A}\bar{A}$. La probabilité de chaque succès est p, la probabilité de chaque échec est $1-p$. En utilisant l'indépendance des répétitions de l'épreuve, la probabilité de cette série est $p^k(1-p)^{n-k}$. Mais la série ci-dessus n'est pas la seule série possible de k succès : il y par exemple k succès suivi de $n-k$ échecs, l'inverse, ou d'autres séries encore. Chacune de ces séries possède une probabilité $p^k(1-p)^{n-k}$ et contribue à la probabilité $P(X = k)$. Au total, on dénombre C_n^k séries de k succès parmi n épreuve, d'où la formule ci-dessus. On note qu'une loi binomiale vérifie bien les conditions pour être une loi de probabilité. Il est évident que $P(X = k) \geq 0$. Par ailleurs :

$$\begin{aligned}\sum_{k=0}^{n} P(X=k) &= \sum_{k=0}^{n} p^k(1-p)^{n-k}C_n^k \\ &= (p+(1-p))^n = 1,\end{aligned}$$

en utilisant la formule du binôme.

La probabilité $P(X = k)$ croît jusqu'à $(n + 1)p$, puis elle décroît. En effet :

$$\begin{aligned}\frac{P(X=k)}{P(X=k-1)} &= \frac{C_n^k p^k (1-p)^{n-k}}{C_n^{k-1} p^{k-1}(1-p)^{n-k+1}} = \frac{C_n^k}{C_n^{k-1}} \frac{p}{1-p} \\ &= \frac{n!}{(n-k)!k!} \frac{(k-1)!(n-k+1)!}{n!} \frac{p}{(1-p)} = \frac{n-k+1}{k} \frac{p}{(1-p)}.\end{aligned}$$

Donc :

$$\begin{aligned}& P(X=k) \geq P(X=k-1) \\ \iff\quad & p(n-k+1) \geq (1-p)k \\ \iff\quad & k \leq (n+1)p.\end{aligned}$$

Exemple Un système de communication à n composantes fonctionne si au moins la moitié de ses composantes fonctionne. On suppose que chaque composante fonctionne indépendamment des autres composantes avec un probabilité p. On note X le nombre, aléatoire, de composantes qui fonctionnent.

– Système à 3 composantes : $X \sim \mathcal{B}(3, p)$. La probabilité que le système fonctionne est :

$$\begin{aligned}P(X \geq 2) &= P(X=2) + P(X=3) \\ &= C_3^2 p^2 (1-p) + C_3^3 p^3 \\ &= p^2(3-2p).\end{aligned}$$

– Système à 5 composantes : $X \sim \mathcal{B}(5, p)$. La probabilité que le système fonctionne est :

$$\begin{aligned}P(X \geq 2) &= P(X=3) + P(X=4) + P(X=5) \\ &= C_5^3 p^3 (1-p)^2 + C_5^4 p^4 (1-p) + C_5^5 p^5 \\ &= 10p^3(3-2p)^2 + 5p^4(1-p) + p^5.\end{aligned}$$

On peut montrer que le second est meilleur que le premier si $p \geq 1/2$ (ce qui est logique, pourquoi ?).

Variable aléatoire de Poisson

Définition 4.3 (variable aléatoire de Poisson) *Une variable aléatoire X est une variable aléatoire de Poisson de paramètre $\lambda > 0$, notée $\mathcal{P}(\lambda)$ si son système de probabilité est*

$$P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad k = 0, 1, \ldots$$

La variable aléatoire de Poisson est un cas limite de la variable aléatoire binomiale. En effet, supposons $X \sim \mathcal{B}(n,p)$ avec $n \to \infty$, $p \to 0$ et $np = \lambda$ reste constant. Alors :

$$\begin{aligned} P(X=k) &= \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= \frac{n!}{k!(n-k)!} (\lambda/n)^k (1-\lambda/n)^{n-k} \\ &= \frac{n(n-1)\dots(n-k+1)}{n^k} \frac{\lambda^k}{k!} (1-\lambda/n)^n (1-\lambda/n)^{-k}. \end{aligned}$$

Le premier et le dernier facteur tendent vers 1 lorsque n tend vers l'infini. D'autre part, $\lim_{n\to\infty}(1-\lambda/n)^n = e^{-\lambda}$. Donc, finalement :

$$\lim_{n\to\infty} P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

En d'autres termes, si on a un grand nombre d'évènements indépendants ayant chacun une probabilité très petite d'occurrence, alors le nombre total d'occurrences est approximativement distribué comme une loi de Poisson. La convergence de la loi binomiale vers la loi de Poisson est illustrée à la figure 2, où le système de probabilité d'une loi binomiale $\mathcal{B}(n,p)$ a été représenté, en conservant le produit $np = 2$ constant et p décroissant de 0,5 à 0,1. On observe que le système de probabilité d'une $\mathcal{B}(20,\ 0,1)$ est très proche de celui d'une $\mathcal{P}(2)$.

La variable aléatoire de Poisson joue un rôle très important en modélisation probabiliste. Il sert à modéliser des phénomènes aléatoires qui ont à chaque répétition une très petite chance de se réaliser, mais dont le nombre de répétitions est très grand. C'est le cas par exemple :

- du nombre d'erreurs typographiques sur une page ;
- du nombre de requêtes sur un serveur par unité de temps (surtout s'il n'est pas très fréquenté ; sinon, il est préférable d'envisager d'autres modèles) ;
- du nombre d'accidents d'avion en un an ;
- du nombre de désintégration α de matériel radioactif par unité de temps ;
- du nombre de mauvaises herbes dans un champ ;
- ...

On remarque la présence du temps ou de l'espace dans beaucoup de ces exemples. En effet, faisons les hypothèses suivantes : 1) la probabilité d'un évènement est très petite dans un très petit intervalle ; 2) les événements dans des intervalles disjoints sont indépendants. Alors, le nombre d'évènements dans un grand intervalle est une variable aléatoire binomiale où p est très petit et n très grand. A la limite, c'est une variable aléatoire de Poisson de paramètre $\lambda = np$.

4.5 Variable aléatoire continue

Une variable aléatoire continue est une variable aléatoire dont la fonction de répartition est continue partout et dérivable presque partout (c'est-à-dire, partout sauf en un nombre de points

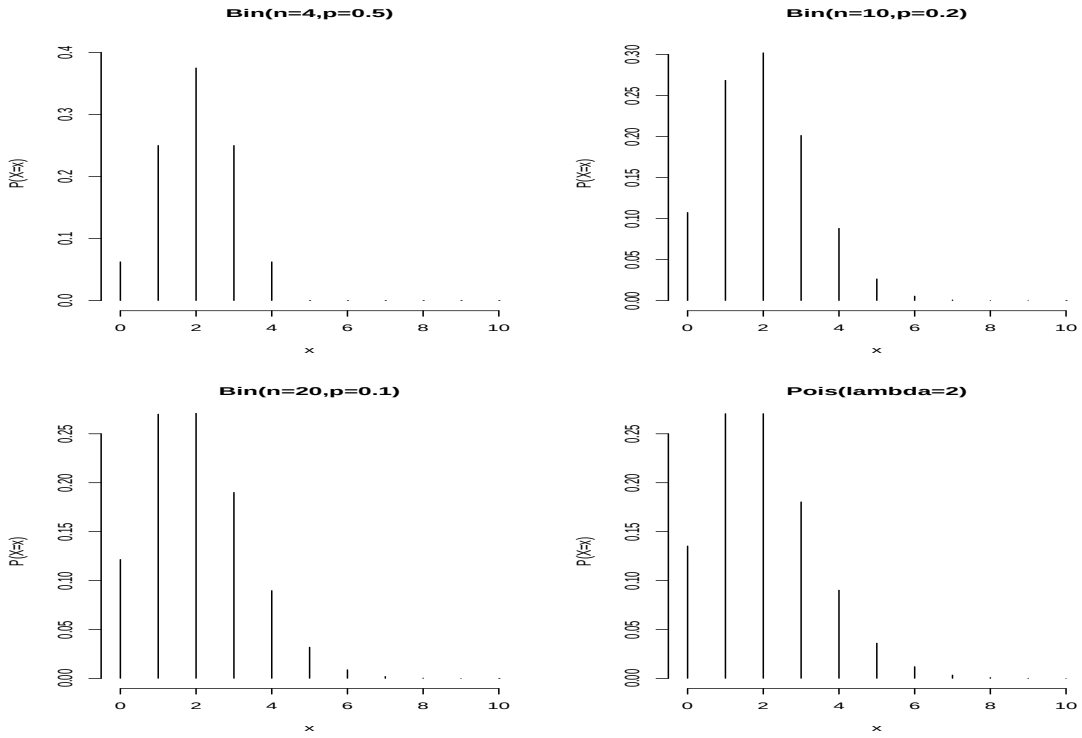

FIG. 2 – Convergence de la loi binomiale vers la loi de Poisson.

au plus dénombrable). Dans ce cas, il existe une fonction $f(x)$ qui permet d'écrire :

$$P(a < X \leq b) = \int_a^b f(x)dx = F(b) - F(a).$$

Aux points où F est dérivable, on a donc $F'(x) = f(x)$. La fonction $f(x)$ s'appelle la **densité de probabilité** de X. On appelle **support** de $f(x)$, la partie de $\mathbf{R}$ où $f(x)$ est non nulle. Les valeurs possibles de la variable aléatoire sont les points où la densité est non nulle. En d'autres termes, l'ensemble des observables $X(\Omega)$ est égal au support de $f(x)$. Contrairement au cas discret, l'ensemble des observables $X(\Omega)$, n'est pas dénombrable.

Voyons maintenant quel sens donner à la densité de probabilité. Comme

$$P(a < X \leq b) = \int_a^b f(x)\,dx,$$

la probabilité que la variable aléatoire X soit comprise entre a et b est l'aire située sous le graphe de la densité de probabilité. Remplaçons b par $a + \delta a$, où δa est très petit. Alors,

$$P(a < X \leq a + \delta a) = \int_a^{a+\delta a} f(x)\,dx \simeq f(a)\delta a.$$

Ainsi, si $f(x)$ n'est pas à proprement parler une probabilité, elle peut toutefois être vue comme une vraisemblance : plus la densité de probabilité est élevée, plus il est vraisemblable que la variable aléatoire prenne des valeurs dans le voisinage de x [5].

Théorème 4.2 (Théorème de caractérisation) *Une fonction $f(x)$ est la densité d'une certaine variable aléatoire continue X (défini sur un certain espace probabilisable $(\Omega, \mathcal{A}, P)$, pas forcément connu) ssi $f(x) \geq 0,\ \forall x \in \mathbf{R}$ et*

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

La première condition assure que les probabilités sont toujours positives. La seconde que

$$P(\Omega) = 1.$$

4.6 Modélisation probabiliste continue

Comme nous l'avons fait pour les variables aléatoires discrètes, nous présentons quelques variables aléatoires qui sont parmi les plus utilisées en modélisation continue.

[5] On peut faire une analogie avec une barre en un matériau dont la densité serait égale à $f(x)$. Si $f(x)$ ne désigne jamais la masse de cette barre en x (qui à strictement parler est nulle), $f(x)dx$ est bien la masse de l'élément dx et $\int_a^b f(x)\,dx$ est la masse de la partie comprise entre a et b.

Variable aléatoire uniforme

Définition 4.4 *Une variable aléatoire X est une variable aléatoire uniforme si sa densité s'écrit :*

$$f(x) = c, \text{ lorsque } a < x < b; \quad f(x) = 0 \text{ sinon.}$$

La constante c ne peut être quelconque. La deuxième condition du théorème de caractérisation entraîne que

$$\int_{-\infty}^{+\infty} c\,dx = 1 \Leftrightarrow c[x]_a^b = 1 \iff c(b-a) = 1,$$

et donc $c = 1/(b-a)$. Si $a = 0$ et $b = 1$, nous retrouvons la loi uniforme standard, où $X \in]0,1[$, qui n'est autre que la fonction RAND sur les calculatrices. La fonction de répartition est $F(x) = 0$ si $x \leq a$ et $F(x) = 1$ si $x \geq b$. Pour $a < x < b$,

$$F(x) = \int_a^x \frac{1}{b-a}\,du = \frac{x-a}{b-a}.$$

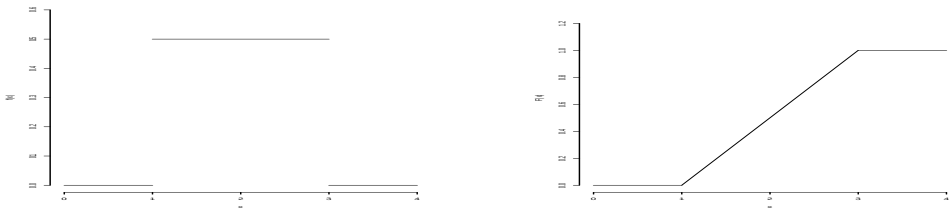

FIG. 3 – Densité (à gauche) et fonction de répartition (à droite) d'une variable aléatoire uniforme sur l'intervalle $[1,3]$.

Variable aléatoire exponentielle

Définition 4.5 *Une variable aléatoire X est une variable aléatoire exponentielle de paramètre $\lambda > 0$ si sa densité s'écrit :*

$$f(x) = \lambda e^{-\lambda x} \text{ lorsque } x > 0; \quad f(x) = 0 \text{ sinon.}$$

La fonction de répartition associée pour $x > 0$ est :

$$F(x) = \int_0^x \lambda e^{-\lambda u}\,du = 1 - e^{-\lambda x},$$

et $F(x) = 0$ pour $x < 0$.

La loi exponentielle est fort utilisée pour modéliser des processus temporels, par exemple :

- la durée de vie de composantes électroniques,
- les temps d'arrivée à un guichet, à un serveur,

FIG. 4 – Densité (à gauche) et fonction de répartition (à droite) d'une variable aléatoire exponentielle de paramètre $\lambda = 2$.

– la décomposition nucléaire, etc...

On notera que l'on retrouve les mêmes exemples que ceux utilisés pour les variables aléatoires de Poisson. En fait, ces deux variables aléatoires sont très liées. On peut montrer que si les temps séparant des évènements sont distribués selon une loi exponentielle, alors le nombre d'évènements par unité de temps est distribué selon une loi de Poisson, et inversement.

La loi exponentielle a une propriété très particulière : l'absence de mémoire du temps déjà attendu. Calculons la probabilité d'attendre un temps supplémentaire s, sachant que l'on a déjà attendu un temps t :

$$\begin{aligned} P(X > s+t \mid X > t) &= \frac{P(\{X > s+t\} \text{ et } \{X > t\})}{P(X > t)} \\ &= \frac{P(X > s+t)}{P(X > t)} \\ &= \frac{1 - P(X \leq s+t)}{1 - P(X \leq t)} \\ &= \frac{1 - F(s+t)}{1 - F(t)} = \frac{1 - (1 - e^{-\lambda(s+t)})}{1 - (1 - e^{-\lambda t})} \\ &= e^{-\lambda s} = 1 - F(s) \\ &= P(X > s), \end{aligned}$$

qui n'est autre que la probabilité d'attendre un temps s, sans attente préalable. Cette propriété indique que **quel que soit** le temps déjà attendu, t, le temps qui reste à attendre est distribué selon la même loi de probabilité. On peut montrer que la densité exponentielle est la seule densité continue qui possède cette propriété. Cette propriété mathématique est-elle en bon accord avec les phénomènes habituellement modélisés à l'aide de la loi exponentielle ? C'est le cas de la désintégration nucléaire. Ce n'est pas trop faux pour les problèmes de guichets et de serveurs. C'est en revanche beaucoup moins vrai pour les données de durée de vie.

Variable aléatoire gaussienne

Définition 4.6 *Une variable aléatoire X est une variable aléatoire gaussienne (ou normale) de paramètres $\mu \in \mathbf{R}$ et $\sigma > 0$, notée $\mathcal{N}(\mu, \sigma^2)$ si sa densité s'écrit :*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad x \in \mathbf{R}.$$

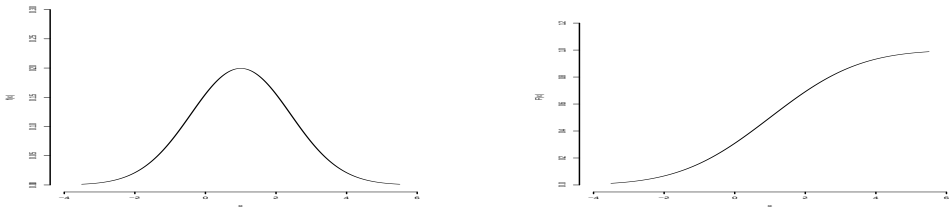

FIG. 5 – Densité (à gauche) et fonction de répartition (à droite) d'une variable aléatoire gaussienne $\mathcal{N}(\mu = 1, \sigma^2 = 4)$.

On peut montrer (mais c'est long) que l'intégrale sur $\mathbf{R}$ de cette densité vaut 1. La densité gaussienne est la densité la plus utilisée en statistique. Il y a pour cela plusieurs raisons qui se renforcent mutuellement.

- Malgré sa forme à première vue complexe, cette densité possède d'excellentes propriétés mathématiques qui font que son utilisation s'avère assez simple[6].
- Comme nous le verrons plus loin, un théorème de convergence (le théorème de la limite centrale) montre que l'addition d'un grand nombre de variables aléatoires indépendantes tend vers une loi gaussienne. En conséquence, cette densité joue un rôle de bassin d'attraction vers lequel les autres densités sont attirées.
- Pour ces deux raisons précédentes, la loi gaussienne est le modèle de bruit universellement adopté en modélisation probabiliste.
- Il se fait que cette densité permet de modéliser assez bien un grand nombre de phénomènes naturels : taille ou poids d'une population, Q.I., etc... Pour expliquer cette constatation expérimentale, on met en avant le théorème de la limite centrale : si le poids, la taille ou le Q.I. peuvent être considérés comme le résultat de la somme d'un grand nombre de causes (ou d'influences) indépendantes, il n'est peut-être pas si étonnant que la variable étudiée soit proche d'une gaussienne.

Un cas particulier important est lorsque $\mu = 0$ et $\sigma = 1$, ce qui définit la variable aléatoire gaussienne standard, notée $\mathcal{N}(0, 1)$, dont la densité est :

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbf{R}.$$

[6] Nous verrons plusieurs de ces propriétés plus loin.

4.7 Fonction d'une variable aléatoire continue

Soit X une variable aléatoire continue dont la densité est f_X. On considère $g(X)$ où g est fonction inversible. Il est évident que si X est une variable aléatoire il en va de même de $Y = g(X)$. On recherche la fonction de répartition F_Y de Y, ainsi que sa densité. Comme g est inversible, c'est une fonction monotone. On considérera que g est croissante[7]. Alors :

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(g(X) \leq y) \\ &= P(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y)). \end{aligned}$$

La densité de Y s'obtient en dérivant F_Y :

$$f_Y(y) = f_X(g^{-1}(y))g'^{-1}(y).$$

Si g est décroissante, les calculs précédents doivent s'adapter pour tenir compte du fait que $g(X) \leq y \iff X \geq g^{-1}(y)$.

Exemple 1 X est une variable aléatoire exponentielle de paramètre λ et $Y = X^2$ (la fonction est inversible car $X \geq 0$). Alors $X = \sqrt{Y}$, et donc

$$F_Y(y) = 1 - e^{-\lambda y^{1/2}} \quad f_Y(y) = \frac{\lambda}{2} y^{-1/2} e^{-\lambda y^{1/2}}, \ y > 0.$$

Exemple 2 Une application importante de cette formule est la suivante. Soit X une gaussienne $\mathcal{N}(0,1)$. Alors

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbf{R}.$$

On définit $Y = \mu + \sigma X$, avec $\sigma > 0$. Alors $x = (y - \mu)/\sigma = g^{-1}(y)$. En appliquant la formule de changement de variable, on trouve

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/2\sigma^2},$$

qui est précisément la densité gaussienne avec les paramètres μ et σ^2.
Ainsi, on vient de montrer un résultat qui sera beaucoup utilisé en statistiques : toute variable aléatoire gaussienne $\mathcal{N}(\mu, \sigma^2)$ se déduit d'une variable aléatoire gaussienne $\mathcal{N}(0,1)$ par

$$\mathcal{N}(\mu, \sigma^2) = \mu + \sigma\mathcal{N}(0,1).$$

[7] Ceci sans perte de généralité, car sinon c'est $-g$ qui est croissante.

5 Couple de variables aléatoires

On s'intéresse maintenant au cas où le résultat d'une expérience aléatoire se décrit à l'aide de deux (ou plusieurs) variables, par exemple :
– le couple poids / taille d'un individu tiré au hasard dans une population ;
– le temps d'arrivée et le temps d'exécution d'une requête sur un serveur ;
– le revenu et le patrimoine d'un ménage tiré au hasard dans une population, etc...
Comme d'habitude, il faudra distinguer le cas discret du cas continu.

5.1 Couple de variables aléatoires discrètes

5.1.1 Loi bivariable

L'outil de description d'un couple de variables aléatoires discrètes, (X, Y), est le système de probabilité bivariable

$$P(X = x, Y = y),$$

défini pour toutes les valeurs x et y prises par les variables aléatoires X et Y respectivement, que l'on notera aussi parfois $p(x, y)$. Ce système de probabilité peut se représenter sous la forme d'un tableau à double entrée :

$y \backslash x$	x_1	x_2	x_3	$\cdots$	x_n	
y_1	$p(x_1, y_1)$	$p(x_2, y_1)$	$p(x_3, y_1)$	$\cdots$	$p(x_n, y_1)$	$p_Y(y_1)$
y_2	$p(x_1, y_2)$	$p(x_2, y_2)$	$p(x_3, y_2)$	$\cdots$	$p(x_n, y_2)$	$p_Y(y_2)$
y_3	$p(x_1, y_3)$	$p(x_2, y_3)$	$p(x_3, y_3)$	$\cdots$	$p(x_n, y_3)$	$p_Y(y_3)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
y_n	$p(x_1, y_n)$	$p(x_2, y_n)$	$p(x_3, y_n)$	$\cdots$	$p(x_n, y_n)$	$p_Y(y_n)$
	$p_X(x_1)$	$p_X(x_2)$	$p_X(x_3)$	$\cdots$	$p_X(x_n)$	1

Il faut bien entendu que la somme $\sum_i \sum_j p(x_i, y_j) = 1$ afin de respecter la condition habituelle sur la somme totale des probabilités.

5.1.2 Loi marginale

Si l'on ne s'intéresse qu'à l'une des deux variables aléatoires uniquement, on utilise la distribution de X seul, que l'on notera

$$p_X(x) = P(X = x).$$

La loi p_X est appelée la loi **marginale** de X. Elle est liée à la distribution du couple de la façon suivante :

$$\begin{aligned} p_X(x) = P(X = x) &= P(X = x, Y \in \mathcal{Y}) \\ &= \sum_{y_j \in \mathcal{Y}} P(X = x, Y = y_j), \end{aligned}$$

en notant $\mathcal{Y}$ l'ensemble des valeurs possibles de la variable aléatoire Y. De même,

$$p_Y(y) = \sum_{x_i \in \mathcal{X}} P(X = x_i, Y = y),$$

où $\mathcal{X}$ est l'ensemble des valeurs possibles de la variable aléatoire X. Le nom de loi marginale provient de ce que dans le tableau bivariable ci-dessus, les probabilités $p_X(x)$ sont obtenues en sommant les colonnes et les probabilités $p_Y(y)$ en sommant les lignes : les résultats sont reportés dans les marges du tableau.

Exemple Une urne contient 3 boules, numérotées 1,2 et 3. On tire successivement 2 boules, sans remise. Soient X et Y le numéro de la premières et deuxième boule, respectivement. On utilise la formule des probabilités composées

$$P(X = x, Y = y) = P(X = x)P(Y = y \mid X = x).$$

On obtient alors le tableau suivant :

$y\backslash x$	1	2	3	
1	0	1/6	1/6	1/3
2	1/6	0	1/6	1/3
3	1/6	1/6	0	1/3
	1/3	1/3	1/3	1

5.2 Couple de variables aléatoires continues

5.2.1 Densité bivariable

Un couple aléatoire (X, Y) est un couple de variables aléatoires continues si chacune des variables aléatoires est continue. On appelle **densité de probabilité** du couple la fonction à deux variables $f(x, y)$ qui possède les propriétés suivantes :

i) $f(x, y) \geq 0$
ii) $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)\, dxdy = 1$
iii) $P((X, Y) \in \mathcal{D}) = \int \int_{\mathcal{D}} f(x, y)dxdy$ où $\mathcal{D}$ est un domaine de $\mathbf{R}^2$.

Une densité bivariable s'interprète de façon similaire à ce qui a été vu au paragraphe 4.5 : la probabilité que le couple (X, Y) soit compris dans le domaine élémentaire $[x, x + dx[\times[y, y + dy[$ est

$$P(X \in [x, x + dx[, Y \in [y, y + dy[) = \int_x^{x+dx} \int_y^{y+dy} f(u, v)dudv \simeq f(x, y)dxdy.$$

Ainsi, plus la densité est élévée en (x, y), plus il est vraisemblable que le couple aléatoire (X, Y) soit autour de (x, y). Il faut remarquer que le couple aléatoire ne peut prendre de valeurs dans la partie de $\mathbf{R}^2$ où la densité est nulle.

Exemple 1 On considère la densité uniforme sur le carré centré en 0, de côté 2. Il faut : a) trouver la densité correspondante, et : b) calculer la probabilité que $(X,Y) \in \mathcal{D}$, où $\mathcal{D}$ est le disque unité.

a) Si la densité est uniforme sur le carré centré en 0, de côté 2, elle est du type $f(x,y) = a$ si $-1 < x < 1$ et $-1 < y < 1$, et $f(x,y) = 0$ ailleurs. Il reste à trouver la constante a. En raison de la propriété ii), l'intégrale de f sur $\mathbf{R}^2$ doit valoir 1. Comme

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy = \int_{-1}^{+1}\int_{-1}^{+1} adxdy = a\int_{-1}^{+1} dx \int_{-1}^{+1} dy = a[x]_{-1}^{1}[y]_{-1}^{1} = 4a,$$

alors $a = 1/4$.

b) $P((X,Y) \in \mathcal{D}) = \frac{1}{4}\int\int_{\mathcal{D}} dxdy$. Cette intégrale est le volume situé sous le graphe de la fonction constante $a = 1/4$, sur un domaine égal au cercle unité. C'est donc l'aire du cercle unité multiplié par 1/4. Donc, $P((X,Y) \in \mathcal{D}) = \pi/4$.

Exemple 2 Soit la densité

$$f(x,y) = \begin{cases} e^{-y} & 0 < x < y < \infty, \\ 0 & \text{ailleurs.} \end{cases}$$

Calculer $P(X \leq 1)$.
Attention ! Cette densité semble ne dépendre que de y, et non de x. En réalité elle dépend de x à travers le domaine sur lequel $f(x,y)$ n'est pas nulle. En effet, $f(x,y)$ est nulle si $y < x$. Le domaine des valeurs possibles pour le couple (X,Y), où $f(x,y) > 0$ est représenté ci-dessous :

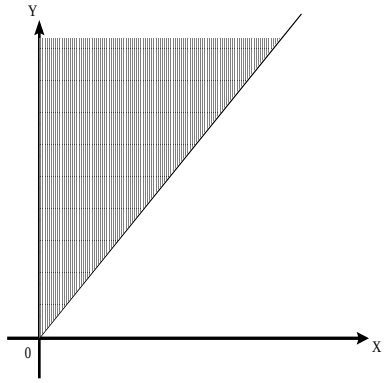

FIG. 6 – Domaine où la densité de probabilité est non nulle.

L'évènement $X \leq 1$ est équivalent à l'évènement $X \leq 1, Y \in \mathbf{R}$. Toutefois, il ne faut pas oublier que la densité est nulle si $y < x$. Il faut donc intégrer la densité sur le domaine où $x \in [0,1]$ et $y \in [x, +\infty]$. Donc,

$$P(X \leq 1) = \int_0^1 \int_x^{+\infty} e^{-y} dydx.$$

Dans cette expression, l'intégrale intérieure porte sur y, et l'intégrale extérieure porte sur x. On commence par calculer l'intégrale intérieure :

$$\int_x^{+\infty} e^{-y} dy = \left[-e^{-y}\right]_x^{+\infty} = 0 - (-e^{-x}) = e^{-x}.$$

En remplaçant l'intégrale intérieure par cette expression, il vient

$$P(X \leq 1) = \int_0^1 e^{-x} dx = \left[-e^{-x}\right]_0^1 = 1 - e^{-1}.$$

5.2.2 Fonction de répartition

La fonction de répartition d'un couple de variables aléatoires est, par définition :

$$\begin{aligned} F(x,y) &= P(X \leq x, Y \leq y) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v) du dv \end{aligned}$$

par application de la propriété iii) du paragraphe précédent. En général, la fonction de répartition bivariable est peu pratique à utiliser, et nous ne la reverrons plus dans ce cours.

5.2.3 Loi marginale

Au paragraphe 5.1 sur les variables aléatoires discrètes, on a vu que lorsqu'on ne s'intéresse qu'à une variable, prise isolément, on parle de loi marginale. Dans le cas continu on définit de même les **densités marginales**.

Définition 5.1 *La densité marginale de X est la fonction définie par :*

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y) dy.$$

Afin d'obtenir la densité marginale de X, on somme la densité bivariable pour toutes les valeurs de y (on retrouve une construction similaire à celle du cas discret, où on sommait toutes les valeurs d'une ligne ou d'une colonne).

Exemple 2 (cont.) Tout au long du chapitre 5, nous reprendrons l'exemple 2 vu ci-dessus :

$$f(x,y) = \begin{cases} e^{-y} & 0 < x < y < \infty, \\ 0 & \text{ailleurs.} \end{cases}$$

Les lois marginales de cette densité sont

$$f_X(x) = \int_x^{+\infty} e^{-y} dy = e^{-x}$$

si $x \geq 0$, et $f_X(x) = 0$ ailleurs,
et,

$$f_Y(y) = \int_0^y e^{-y} dx = e^{-y} [x]_0^y = ye^{-y}$$

si $y \geq 0$ et $f_Y(y) = 0$ ailleurs. Il faut noter que dans cet exemple, les domaines où les densités marginales sont non nulles est $\mathbf{R}^+$.

5.3 Indépendance

Nous avions vu au paragraphe 3.2 que deux événements A et B sont indépendants lorsque $P(A \cap B) = P(A)P(B)$. Cette définition s'étend aux variables aléatoires de la façon suivante :

Définition 5.2 *a) Deux variables aléatoires discrètes sont indépendantes ssi*

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

pour tous x et y.
b) Deux variables aléatoires continues sont indépendantes ssi

$$f(x, y) = f_X(x) f_Y(y)$$

pour tous $(x, y) \in \mathbf{R}^2$.

Remarques
a) est une conséquence immédiate de

$$P(A \cap B) = P(A)P(B) \ ;$$

b) peut se comprendre de deux façons :

1. Soient les évènements $A = \{X \leq x\}$ et $B = \{Y \leq y\}$. Alors

   $$P(A \cap B) = P(X \leq x, Y \leq y).$$

   Si X et Y sont indépendants, il faut que $P(A \cap B) = P(A)P(B)$ pour tout x et tout y. Donc, il faut que

   $$F(x, y) = F_X(x).F_Y(y).$$

   En dérivant par x puis par y cette dernière expression, il vient

   $$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y} = \frac{\partial^2 F_X(x) F_Y(y)}{\partial x \partial y} = f_X(x) f_Y(y).$$

2. Soient les évènements $C = \{x < X < x + dx\}$ et $D = \{y < Y < y + dy\}$. Alors $P(A \cap B) = P(A)P(B)$ s'écrit

   $$f(x, y)dxdy = f_X(x)dx f_Y(y)dy.$$

Exemple 2 (cont.) Le produit des densités marginales $f_X(x)f_Y(y) = e^{-x}e^{-y}$, pour $(x,y) \in \mathbf{R}^2$, est différent de la densité bivariable. Donc, dans cet exemple, les variables aléatoires X et Y ne sont pas indépendantes.

Théorème 5.1 *Pour que deux variables aléatoires continues X et Y soient indépendantes, il faut et il suffit que la densité soit factorisable de la manière suivante :*

$$f(x,y) = g(x)h(y) \qquad \forall (x,y) \in \mathbf{R}^2.$$

Dans ce cas, il existe une constante a telle que $f_X(x) = ag(x)$ et $f_Y(y) = h(y)/a$.

Preuve :
a) La condition est nécessaire : c'est trivial avec $g(x) = f_X(x)$ et $h(y) = f_Y(y)$.
b) La condition est suffisante : on suppose que la propriété de factorisation est vérifiée. Alors

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = g(x) \int_{-\infty}^{\infty} h(y)dy.$$

Dans ce cas, on pose $a = \int_{-\infty}^{\infty} h(y)dy$. De même,

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = h(y) \int_{-\infty}^{\infty} g(x)dx.$$

Comme

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = \int_{-\infty}^{\infty} g(x)dx \int_{-\infty}^{\infty} h(y)dy = 1,$$

Alors, $\int_{-\infty}^{\infty} g(x)dx = 1/a$. □

Un corollaire important de ce théorème est que le domaine B sur lequel la densité $f(x,y)$ est non nulle doit être factorisable, ce qui signifie qu'il doit s'écrire comme le produit cartésien des domaines B_X et B_Y sur lesquels les densités marginales $f_X(x)$ et $f_Y(y)$ sont non nulles. Ce qui s'énonce sous la forme du théorème suivant :

Théorème 5.2 *Pour que les variables aléatoires X et Y soient indépendantes, il est nécessaire (mais non suffisant) que*

$$B = B_X \times B_Y.$$

Exemple 2 (cont.) On reprend la densité habituelle, $f(x,y) = e^{-y}$ $0 < x < y < \infty$, et $f(x,y) = 0$ sinon. Le domaine B est le secteur de $\mathbf{R}^{+2}$ délimité par la première bissectrice $y = x$ et par l'axe des ordonnées $x = 0$. Ce domaine ne peut pas se factoriser sous la forme du produit cartésien de deux intervalles. Par conséquent, les variables aléatoires ne peuvent pas être indépendantes.

Exemple 3 Soit la densité bivariable $f(x,y) = (1+x+y)/2$ $0 < x,y < 1$, et $f(x,y) = 0$ sinon. Le domaine B est le rectangle $]0,1[\times]0,1[$, mais la densité ne se factorise pas. Donc, les variables

aléatoires X et Y ne sont pas indépendantes.

Exemple 4 Soit la densité bivariable $f(x,y) = 4xy \;\; 0 < x, y < 1$, et $f(x,y) = 0$ sinon. Le domaine B est le rectangle $]0,1[\times]0,1[$. De plus, la densité se factorise. Donc, les variables aléatoires X et Y sont indépendantes.

On termine le paragraphe sur l'indépendance avec le théorème (non démontré) suivant :

Théorème 5.3 *Soit X et Y deux variables aléatoires indépendantes. Alors $u(X)$ et $v(Y)$ sont également des variables aléatoires indépendantes, quelles que soient les fonctions u et v.*

Exemple 5 Si X et Y sont deux variables aléatoires gaussiennes indépendantes, alors X^2 et Y^2 sont également indépendantes.

5.4 Distributions conditionnelles

On considère toujours un couple de variables aléatoires (X,Y). Contrairement à la distribution marginale $f_X(x)$, où Y n'est pas précisé et peut donc prendre toutes les valeurs, on va maintenant fixer la valeur de Y (resp. de X), et voir comment évolue la distribution de X (resp. de Y).

5.4.1 Cas discret

Dans le cas discret, la distribution conditionnelle se définit facilement à l'aide des probabilités conditionnelles sur les évènements :

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}.$$

Au dénominateur apparaît la probabilité marginale de Y.

Si X et Y sont indépendants, alors :

$$\begin{aligned} P(X = x \mid Y = y) &= \frac{P(X = x, Y = y)}{P(Y = y)} \\ &= \frac{P(X = x)P(Y = y)}{P(Y = y)} \\ &= P(X = x), \end{aligned}$$

et dans ce cas, la distribution conditionnelle est égale à la distribution marginale.

5.4.2 Cas continu

Pour passer au cas continu, on remplace comme d'habitude les probabilités par des densités.

Définition 5.3 *La densité conditionnelle de X, sachant $Y = y$ est*

$$f_{X|Y}(x \mid Y = y) = \frac{f(x,y)}{f_Y(y)}$$

sous réserve que $f_Y(y) > 0$. Si $f_Y(y) = 0$, alors $f_{X|Y}(x \mid Y = y) = 0$.

Exemple 2 (cont.) Comme $f_X(x) = e^{-x}$, $x > 0$ et $f_Y(y) = ye^{-y}$, $y > 0$, la densité conditionnelle de X, sachant $Y = y$ est :

$$f_{X|Y}(x \mid Y = y) = \frac{e^{-y}}{ye^{-y}} = \frac{1}{y},\ 0 < x < y.$$

et $f_{X|Y}(x \mid Y = y) = 0$ sinon. Donc, conditionnellement à $Y = y$, X est une variable aléatoire uniforme sur $]0, y[$. La densité conditionnelle de Y, sachant$X = x$ est :

$$f_{Y|X}(y \mid X = x) = \frac{e^{-y}}{e^{-x}} = e^{-(y-x)},\ 0 < x < y.$$

et$f_{Y|X}(y \mid X = x) = 0$ sinon. Donc, conditionnellement à $X = x$, $Y - x$ est une variable aléatoire exponentielle de paramètre 1.

Si X et Y sont des variables aléatoires indépendantes, on a

$$f_{X|Y}(x \mid Y = y) = \frac{f(x,y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x),$$

et dans ce cas, la distribution conditionnelle est égale à la distribution marginale.

Bien entendu, la relation

$$f_{X|Y}(x \mid Y = y) = \frac{f(x,y)}{f_Y(y)}$$

peut s'inverser pour donner

$$f(x,y) = f_{X|Y}(x \mid Y = y)f_Y(y).$$

Cette relation permet de construire des densités bivariables à partir d'une densité marginale et d'une densité conditionnelle.

Exemple 6 On considère l'expérience suivante : on tire au hasard un nombre X entre 0 et 1, puis un nombre au hasard entre 0 et X. On formalise cela de la façon suivante : X est une variable aléatoire uniforme sur $]0, 1[$. Conditionnellement à $X = x$, Y est une variable aléatoire uniforme

sur $]0, x[$, c'est-à-dire $f_X(x) = 1,\ 0 < x < 1$ et $f_{Y|X}(y \mid X = x) = 1/x,\ 0 < y < x$.
La loi bivariable est

$$f(x, y) = f_{Y|X}(y \mid X = x) f_X(x) = \frac{1}{x}, \quad 0 < y < x < 1.$$

La loi marginale de Y est

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_y^1 \frac{1}{x} dx = [\ln x]_y^1 = -\ln y, \quad 0 < y < 1.$$

La densité conditionnelle de X sachant $Y = y$ est donc

$$f_{X|Y}(x \mid Y = y) = -\frac{1}{x \ln y}, \quad 0, y < x < 1,$$

et $f_{X|Y}(x \mid Y = y) = 0$ ailleurs.

6 Sommes de variables aléatoires indépendantes

6.1 Cas discret

Soient X et Y deux variables aléatoires discrètes indépendantes. On recherche la distribution de la nouvelle variable aléatoire $Z = X + Y$:

$$\begin{aligned} P(Z=z) &= P(X+Y=z) \\ &= \sum_{x\in\mathcal{X}} P(X=x, X+Y=z) \\ &= \sum_{x\in\mathcal{X}} P(X=x, Y=z-x) \\ &= \sum_{x\in\mathcal{X}} P(X=x)P(Y=z-x), \end{aligned}$$

où $\mathcal{X}$ désigne comme d'habitude l'ensemble de valeurs possibles pour la variable aléatoire X. Le passage de la ligne 1 à 2 est une application de la formule des probabilités totales et composées, celui de la ligne 3 à la ligne 4 se justifie par l'indépendance des variables aléatoires X et Y.

Exemple 1 **[Variables aléatoires binomiales]** Soit X une variable aléatoire de Bernoulli de paramètre p et Y une variable aléatoire binomiale de paramètres (n, p), indépendante de X. On montre que $Z = X + Y$ est une variable aléatoire binomiale de paramètres $(n + 1, p)$. En effet,

$$\begin{aligned} P(Z=z) &= P(X=0, Y=z) + P(X=1, Y=z-1) \\ &= (1-p)\, C_n^z p^z (1-p)^{n-z} + p\, C_n^{z-1} p^{z-1} (1-p)^{n+1-z} \\ &= p^z (1-p)^{n+1-z} \left(C_n^z + C_n^{z-1} \right) \\ &= p^z (1-p)^{n+1-z}\, C_{n+1}^z. \end{aligned}$$

On reconnaît pour Z une loi binomiale de paramètres $(n+1, p)$. La première ligne est l'application directe de la formule générale. On applique ensuite la définition de la loi binomiale. Le passage à la dernière ligne avait été démontré au paragraphe 2.3

Un premier corollaire de cet exemple est qu'une variable aléatoire binomiale de paramètres (n, p) peut être vue comme la somme de n variables aléatoires de Bernoulli indépendantes, de même paramètre p. Pour s'en convaincre, il suffit de raisonner par récurrence, et voir qu'une variable aléatoire de Bernoulli est aussi une variable aléatoire binomiale de paramètres $(1, p)$.

Un second corollaire est que si X est une variable aléatoire binomiale de paramètres (n, p) et Y une variable aléatoire binomiale de paramètres (m, p) indépendante de X, alors $Z = X + Y$ est une variable aléatoire binomiale de paramètres $(n + m, p)$, ce qui s'écrit symboliquement

$$\mathcal{B}(n, p) + \mathcal{B}(m, p) \sim \mathcal{B}(n + m, p),$$

où le signe $\sim$ signifie « est distribué comme ».

Pour démontrer ce résultat, il suffit de voir que chaque variable aléatoire binomiale peut s'écrire comme la somme de variables aléatoires de Bernoulli indépendantes.

Exemple 2 **[Variables aléatoires de Poisson]** Soient X et Y deux variables aléatoires de Poisson indépendantes, de paramètres λ et μ respectivement. On montre que $Z = X + Y$ est une variable aléatoire de Poisson de paramètre $\lambda + \mu$. En effet,

$$\begin{aligned} P(Z = n) &= \sum_{x=0}^{n} P(X = x)P(Y = n - x) \\ &= \sum_{x=0}^{n} e^{-\lambda}\frac{\lambda^x}{x!}\, e^{-\mu}\frac{\mu^{n-x}}{(n-x)!} \\ &= e^{-(\lambda+\mu)} \sum_{x=0}^{n} \lambda^x \mu^{n-x} \frac{1}{x!(n-x)!} \\ &= \frac{e^{-(\lambda+\mu)}}{n!} \sum_{x=0}^{n} \frac{n!}{x!(n-x)!} \lambda^x \mu^{n-x} \\ &= \frac{e^{-(\lambda+\mu)}(\lambda+\mu)^n}{n!}. \end{aligned}$$

A la quatrième ligne, on a simplement divisé et multiplié par $n!$. La dernière ligne s'obtient par l'application de la formule du binôme. Ce résultat s'écrit symboliquement :

$$\mathcal{P}(\lambda) + \mathcal{P}(\mu) \sim \mathcal{P}(\lambda + \mu).$$

6.2 Cas continu

Selon le principe habituel, on adapte la formule précédente au cas continu en changeant les probabilités par les densités et les sommes discrètes par des intégrales. Si X et Y sont des variables aléatoires continues indépendantes, la densité de $Z = X + Y$ est liée aux densités de X et Y par la formule suivante :

$$f_Z(z) = \int_{\mathbf{R}} f_X(x) f_Y(z - x)\, dx.$$

Cette opération s'appelle le produit de convolution de f_X par f_Y.

Exemple 3 **[Variables aléatoires uniformes]** X et Y sont des variables aléatoires indépendantes, uniformes sur $[0, 1]$. La variable aléatoire Z est donc nécessairement comprise entre 0 et 2. Il en résulte que si $z < 0$ ou si $z > 2$, la densité $f_Z(z)$ est nulle. On distingue deux cas lorsque la densité n'est pas nulle :

– si $0 < z < 1$: la densité $f_Y(z - x)$ est nulle si $z - x < 0$, donc les bornes de l'intégrale vont de 0 à z, et

$$f_Z(z) = \int_0^z 1.1\, dy = z.$$

– si $1 < z < 2$, la densité $f_Y(z-x)$ est nulle si $z - x > 1$, donc les bornes de l'intégrale vont de $z-1$ à 1, et

$$f_Z(z) = \int_{z-1}^{1} 1.1\, dy = 2 - z.$$

Finalement,

$$f_Z(z) = \begin{cases} z & \text{si } 0 \leq z < 1, \\ 2-z & \text{si } 1 \leq z < 2, \\ 0 & \text{sinon.} \end{cases}$$

En raison de la forme du graphe de cette densité, elle est parfois appelée *densité triangulaire*.

Exemple 3 **[Variables aléatoires gaussiennes]** Soient X et Y deux variables aléatoires gaussiennes $\mathcal{N}(0,1)$ indépendantes. L'application de la formule donne

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} e^{-x^2/2} e^{-(z-x)^2/2}\, dx.$$

L'argument de l'exponentielle est (au signe près)

$$\begin{aligned} x^2/2 + (z-x)^2/2 &= (x^2 + z^2 - 2xz + x^2)/2 \\ &= (x - z/2)^2 + z^2/4. \end{aligned}$$

En divisant et multipliant par $1/\sqrt{2}$ le facteur devant l'intégrale, la densité devient :

$$f_Z(z) = \frac{1}{\sqrt{2\pi}1/\sqrt{2}} \int_{\mathbf{R}} e^{-(x-z/2)^2/(2.1/2)}\, dx \,.\, \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-z^2/4}.$$

Le premier facteur est l'intégrale de la densité d'une gaussienne $\mathcal{N}(z/2, 1/2)$, qui vaut donc 1. Il reste

$$f_Z(z) = \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-z^2/4}$$

qui est la densité d'une gaussienne $\mathcal{N}(0,2)$.

Ainsi, on vient de démontrer que

$$\mathcal{N}(0,1) + \mathcal{N}(0,1) \sim \mathcal{N}(0,2).$$

En procédant de la même manière, on peut également montrer que

$$\mathcal{N}(\mu_1, \sigma_1^2) + \mathcal{N}(\mu_2, \sigma_2^2) \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2, +\sigma_2^2),$$

ce qui en français se dit :

La somme de deux variables aléatoires gaussiennes indépendantes est une variable aléatoire gaussienne dont l'espérance est la somme des espérances et la variance est la somme des variances.

7 Espérance, variance et covariance

7.1 L'espérance mathématique

L'espérance mathématique d'une variable aléatoire X, notée $E[X]$, est un **paramètre de position** qui se définit de la façon suivante :

Définition 7.1 *a) Si X est une variable aléatoire discrète, alors*

$$E[X] = \sum_{x \in \mathcal{X}} xP(X = x).$$

b) Si X est une variable aléatoire continue, alors

$$E[X] = \int_{\mathbf{R}} xf(x)dx.$$

sous réserve que ces sommes existent et soient finies[8].

L'espérance mathématique peut être vue comme le barycentre de la loi de probabilité. La figure 7 en est une illustration. A gauche la distribution d'une variable aléatoire discrète est représentée à l'aide d'un diagramme en bâtons, dont il faut imaginer qu'ils sont pesants. A droite est représentée une densité de probabilité, qui représente par exemple la manière dont un matériau pesant est posé sur l'axe des abscisses. Dans les deux cas, l'espérance mathématique est le point d'équilibre de l'axe des abscisses (indiqué par une ligne verticale).

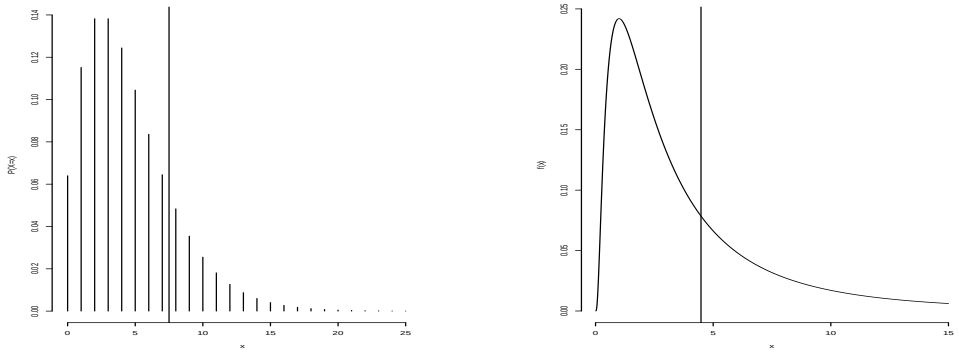

FIG. 7 – Espérance mathématique d'une variable aléatoire discrète (à gauche) et continue (à droite).

Nous allons maintenant calculer l'espérance mathématique des lois vues au chapitre 4.

Variable aléatoire de Bernoulli L'application de la formule donne

$$E[X] = 0.P(X = 0) + 1.P(X = 1) = p.$$

[8] Ce qui est assuré si les conditions $\sum_{x \in \mathcal{X}} |x|P(X = x) < \infty$ et $\int_{\mathbf{R}} |x|f(x)dx < \infty$ sont vérifiées.

Variable aléatoire binomiale On montre que si $X \sim \mathcal{B}(n,p)$, alors $E[X] = np$. En effet :

$$\begin{aligned} E[X] &= \sum_{x=0}^{n} x\, p^x q^{n-x} \frac{n!}{x!(n-x)!} \\ &= 0 + np \sum_{x=1}^{n} p^{x-1} q^{(n-1)-(x-1)} \frac{(n-1)!}{(x-1)!((n-1)-(x-1))!} \\ &= np \sum_{y=0}^{n-1} p^y q^{(n-1)-y} \frac{(n-1)!}{y!((n-1)-y)!} \\ &= np. \end{aligned}$$

La seconde égalité résulte de la mise en facteur de np et de la simplification de la fraction par x. Pour la troisième égalité, on a fait le changement de variable $y = x - 1$. La somme vaut alors 1, par application du théorème du binôme.
Nous verrons plus loin (paragraphe 7.5) une démonstration beaucoup plus simple de ce résultat.

Variable aléatoire de Poisson Si $X \sim \mathcal{P}(\lambda)$, alors $E[X] = \lambda$. En effet :

$$\begin{aligned} E[X] &= \sum_{x=0}^{\infty} x e^{-\lambda} \frac{\lambda^x}{x!} \\ &= 0 + \lambda \sum_{x=1}^{\infty} e^{-\lambda} \frac{\lambda^{x-1}}{(x-1)!} \\ &= \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \\ &= \lambda e^{-\lambda} e^{\lambda} = \lambda. \end{aligned}$$

La technique est similaire à celle utilisée précédemment. A la troisième ligne on reconnaît le développement en série infinie de e^λ, ce qui permet de conclure.
On aurait pu obtenir ce résultat en observant que si une loi de Poisson est la limite d'une loi binomiale pour $p \to 0$, $n \to \infty$, $np = \lambda$, alors l'espérance de cette loi doit être égale à $np = \lambda$.

Variable aléatoire géométrique Si X est distribuée selon une loi géométrique de paramètre $p > 0$, alors $E[X] = 1/p$. En effet, si on note $q = 1 - p$,

$$\begin{aligned} E[X] &= \sum_{x=1}^{\infty} x\, pq^{x-1} \\ &= p \sum_{x=1}^{\infty} x\, q^{x-1} \\ &= p \left(\sum_{x=1}^{\infty} q^x \right)'. \end{aligned}$$

où $(\cdot)'$ désigne la dérivée par rapport à q. Un résultat classique des séries donne que $\sum_{x=1}^{\infty} q^x = q/(1-q)$. Comme d'autre part $(q/(1-q))' = 1/(1-q)^2$, on a finalement

$$E[X] = p/(1-q)^2 = p/p^2 = 1/p.$$

Ce résultat est somme toute assez logique. Si un évènement a une probabilité p de se produire, il faut attendre, en espérance, $1/p$ essais avant de voir le premier succès. Ainsi, aux dés, il faut lancer en espérance 6 fois le dé pour voir une face particulière apparaître pour la première fois.

Variable aléatoire uniforme Si X est une variable aléatoire uniforme sur $]a,b[$, alors $E[X] = (a+b)/2$. En effet,

$$E[X] = \int_a^b x \frac{1}{b-a} dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{1}{2}\frac{b^2-a^2}{b-a} = \frac{a+b}{2}.$$

Ce résultat n'est pas surprenant : lorsque la densité est uniforme sur un intervalle, donc constante sur cet intervalle, l'espérance se trouve au centre de l'intervalle.

Variable aléatoire exponentielle Si X est une variable aléatoire exponentielle de paramètre $\lambda > 0$, alors $E[X] = 1/\lambda$. Le calcul de l'espérance se fait par intégration par partie :

$$\begin{aligned} E[X] &= \lambda \int_0^\infty x e^{-\lambda x}\, dx \\ &= -\lambda \left[\frac{1}{\lambda} x e^{-\lambda x}\right]_0^\infty + \lambda \int_0^\infty \frac{1}{\lambda} e^{-\lambda x} dx \\ &= -0 + 0 + \int_0^\infty e^{-\lambda x} dx \\ &= -\frac{1}{\lambda}\left[e^{-\lambda x}\right]_0^\infty \\ &= \frac{1}{\lambda}. \end{aligned}$$

Variable aléatoire gaussienne Si X est une variable aléatoire gaussienne $\mathcal{N}(0,1)$, alors $E[X] = 0$. En effet,

$$E[X] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty x\, e^{-x^2/2} = 0,$$

car il s'agit de l'intégrale d'une fonction impaire.

7.2 Espérance d'une fonction d'une variable aléatoire

Théorème 7.1 *Soit X une variable aléatoire d'espérance mathématique finie. Alors*

$$E[aX + b] = aE[X] + b.$$

Preuve : On démontre le résultat pour une variable aléatoire continue de densité $f(x)$. La démonstration pour une variable aléatoire discrète est laissée à titre d'exercice.

$$\begin{aligned} E[aX+b] &= \int_{-\infty}^\infty (ax+b) f(x) dx \\ &= a\int_{-\infty}^\infty x f(x) dx + b\int_{-\infty}^\infty f(x) dx \\ &= aE[X] + b. \end{aligned}$$

□

Notons deux cas particuliers de ce théorème :

- Si $a = 0$, alors $E[b] = b$: l'espérance mathématique d'une constante (donc non aléatoire) est cette constante.
- $E[\,X - E[X]\,] = E[X] - E[X] = 0$: lorsque l'on retranche à une variable aléatoire son espérance mathématique, la variable aléatoire ainsi obtenue a une espérance mathématique nulle. Cette opération s'appelle *centrer* une variable aléatoire.

Exemple Soit X une variable aléatoire gaussienne $\mathcal{N}(0,1)$, et soit Y une gaussienne $\mathcal{N}(\mu, \sigma^2)$. On a vu précédemment que d'une part $E[X] = 0$ et d'autre part $Y = \mu + \sigma X$. Par conséquent,

$$E[Y] = E[\mu + \sigma X] = \mu + \sigma E[X] = \mu.$$

Théorème 7.2 *Soit X une variable aléatoire (discrète ou continue), g une fonction quelconque et $Y = g(X)$. Alors,*

$$E[Y] = E[g(X)] = \begin{cases} \sum_x g(x) P(X = x) & \text{(variable aléatoire discrète);} \\ \\ \int_{-\infty}^{\infty} g(x) f(x) dx & \text{(variable aléatoire continue).} \end{cases}$$

Preuve : On présente une démonstration simplifiée qui suppose que g est bijective. Une démonstration plus complète, qui ne fait pas cette hypothèse existe, mais sort du cadre de ce cours.

- Cas discret :

$$E[Y] = \sum_y y P(Y = y) = \sum_x g(x) P(Y = y),$$

car $y = g(x)$. D'autre part,

$$Y = y \iff g(X) = g(x) \iff X = x.$$

Donc, $E[Y] = \sum_x g(x) P(X = x)$.

- Cas continu (on présente une démonstration pour le cas où g est croissante). Rappelons que dans ce cas, la densité de $g(X)$ est $f_Y(y) = f_X(g^{-1}(y)) g^{-1}(y)'$. Alors

$$E[g(X)] = E[Y] = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_{-\infty}^{\infty} y f_X(g^{-1}(y)) g^{-1}(y)' dy.$$

On fait le changement de variable $y = g(x) \iff x = g^{-1}(y) \iff dx = g^{-1}(y)' dy$, ce qui mène à

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx.$$

□

7.3 La variance

Définition 7.2 *Pour une variable aléatoire X dont $E[X^2] < \infty$, la variance est*

$$Var(X) = E[\,(X - E[X])^2\,].$$

Nous avions vu que l'espérance mathématique est un paramètre de position de la variable aléatoire. La variance est un **paramètre de dispersion** de la variable aléatoire autour de son espérance mathématique. On dit aussi que c'est un paramètre de variabilité. Elle indique de combien (en espérance) la variable aléatoire s'écarte de l'espérance. Pourquoi cette dispersion se définit-elle à partir d'un écart quadratique (quadratique signifie « au carré») ? Si on cherche à calculer une dispersion « moyenne », on considérera de façon naturelle l'écart entre X et son espérance. L'espérance mathématique de cet écart est nulle, nous l'avons vu au paragraphe précédent (car, par construction, les valeurs supérieures à l'espérance compensent celles inférieures à l'espérance). On pourrait calculer $E[\,|X-E[X]|\,]$ car la valeur absolue a l'avantage d'être toujours positive. L'inconvénient est que cette grandeur ne possède pas de bonnes propriétés mathématiques, notamment parce que la fonction $|x|$ n'est pas dérivable en 0. Finalement, $E[\,(X-E[X])^2\,]$ a l'avantage de la valeur absolue, sans son inconvénient. En fait, nous aurons l'occasion de voir que la fonction quadratique possède l'immense avantage de mener à des équations linéaires lorsqu'on la minimise. C'est la raison principale de son emploi quasi universel.

La variance possède les propriétés suivantes :

Propriété 7.1 *Soit X une variable aléatoire de variance finie et soit a, b deux réels finis. Alors,*

$$Var(aX + b) = a^2 Var(X).$$

Preuve :

$$\begin{aligned} Var(aX+b) &= E[\,(aX+b-E[aX+b])^2\,] \\ &= E[\,(aX+b-aE[X]-b)^2\,] \\ &= a^2 E[\,(X-E[X])^2\,] \\ &= a^2 Var(X). \end{aligned}$$

□

Deux conséquences de ce résultat méritent un commentaire.

1. D'une part on voit que $Var(X+b) = Var(X)$: translater une variable aléatoire n'affecte pas sa variance. Il s'agit donc bien d'une grandeur qui n'est pas liée à la position. Elle "dit autre chose" que l'espérance.

2. D'autre part, $Var(aX) = a^2 Var(X)$: la variance ne s'exprime pas dans l'unité de mesure de la variable X, mais selon le carré de cette unité. Ainsi, si X est une durée exprimée en secondes (par exemple la durée de vie d'un composant électronique), alors $Var(X)$ sera en s^2. Si X est une taille exprimée en cm, $Var(X)$ sera exprimée en cm^2, mais ne sera pas une surface ! Afin de retrouver l'unité d'origine, on prend la racine carrée de la variance. C'est ce qu'on introduit dans la définition suivante.

Définition 7.3 *Soit X une variable aléatoire dont la variance est finie, On appelle écart-type de X la grandeur $\sigma = \sqrt{Var(X)}$.*

Propriété 7.2 *Il existe une formule qui simplifie le calcul de la variance d'une variable aléatoire. Soit X une variable aléatoire de variance finie. Alors,*

$$Var(X) = E[X^2] - E[X]^2.$$

Preuve

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

□

Nous allons calculer maintenant les variances des variable aléatoires habituelles :

Variable aléatoire de Bernoulli Comme $E[X^2] = 0^2 P(X = 0) + 1^2 P(X = 1) = p$, il vient

$$Var(X) = p - p^2 = p(1 - p).$$

Variable aléatoire binomiale On peut montrer que si $X \sim \mathcal{B}(n, p)$, alors

$$Var(X) = np(1 - p).$$

La démonstration directe est longue et technique. On verra au paragraphe 7.5 une démonstration indirecte mais rapide de ce résultat.

Variable aléatoire géométrique Si X est une variable aléatoire géométrique de paramètre p, alors $Var(X) = (1-p)/p^2$. Ici encore, la démonstration directe est longue et inintéressante. Une démonstration assez brève existe, mais utilise des notions qui ne sont pas définies dans ce cours.

Variable aléatoire de Poisson Si $X \sim \mathcal{P}(\lambda)$, alors

$$Var(X) = \lambda.$$

La démonstration de ce résultat utilise des notions qui se situent hors du cadre de ce cours. Néanmoins, on peut observer que si on considère une $\mathcal{P}(\lambda)$ comme la limite de $Y_n \sim \mathcal{B}(n,p)$ lorsque $n \to \infty$, $p \to 0$ avec $np = \lambda$, alors on doit avoir que $Var(Y_n) \to \lambda = Var(X)$.

Variable aléatoire uniforme

$$\begin{aligned} E[X^2] &= \frac{1}{b-a}\int_a^b x^2\,dx \\ &= \frac{1}{3(b-a)}\left[x^3\right]_a^b \\ &= \frac{b^2+ab+a^2}{3}. \end{aligned}$$

Donc,

$$\begin{aligned} Var(X) &= E[X^2] - E[X]^2 \\ &= \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} \\ &= \frac{(a-b)^2}{12}. \end{aligned}$$

Variable aléatoire exponentielle On rappelle que pour une variable aléatoire exponentielle, $E[X] = 1/\lambda$. L'espérance de X^2 se calcule à l'aide d'une intégrale par partie :

$$\begin{aligned} E[X^2] &= \int_0^\infty \lambda x^2 e^{-\lambda x}\,dx \\ &= \left[-\lambda x^2 \frac{1}{\lambda} e^{-\lambda x}\right]_0^\infty - \int_0^\infty -2\lambda x \frac{1}{\lambda} e^{-\lambda x}\,dx \\ &= \frac{2}{\lambda}\int_0^\infty \lambda x e^{-\lambda x}\,dx \\ &= \frac{2}{\lambda}E[X] \\ &= \frac{2}{\lambda^2}. \end{aligned}$$

Donc,

$$Var(X) = E[X^2] - E[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Variable aléatoire gaussienne On considère d'abord une variable aléatoire gaussienne $Y \sim \mathcal{N}(0,1)$. Alors $E[Y] = 0$. On calcule $E[Y^2]$ par partie :

$$E[X^2] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty x^2 e^{-x^2/2}\,dx$$

$$\begin{aligned} &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x.xe^{-x^2/2}\, dx \\ &= \left[-\frac{x}{\sqrt{2\pi}} e^{-x^2/2}\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} -\frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \\ &= 0 + 1, \end{aligned}$$

car le premier terme est nul (en effet, l'exponentielle négative, qui tend vers 0, l'emporte sur tous les polynômes de degré fini) et le second terme est la somme de la densité gaussienne, donc vaut 1. D'où

$$Var(X) = 1 - E[X]^2 = 1 - 0 = 1.$$

On considère maintenant une variable aléatoire gaussienne $X \sim \mathcal{N}(\mu, \sigma^2)$. Alors, nous avons vu que $X = \mu + \sigma Y$. Par application de la propriété 7.1,

$$E[X] = \mu \text{ et } Var(X) = \sigma^2.$$

Ainsi, on voit maintenant qu'une variable aléatoire gaussienne est paramétrée par son espérance μ et sa variance σ^2.

7.4 La covariance

Dans tout ce paragraphe, on supposera que les variables aléatoires X et Y ont une espérance et une variance finies et que $E[XY] < \infty$.

Définition 7.4 *La covariance entre X et Y, notée $Cov(X, Y)$ est la grandeur*

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])].$$

Propriété 7.3 *La covariance possède les propriétés suivantes :*

1. $Cov(X, Y) = Cov(Y, X)$
2. $Cov(X, X) = Var(X)$
3. $Cov(X, Y) = E[XY] - E[X]E[Y]$
4. $Cov(aX + b, cY + d) = acCov(X, Y)$

Preuve : Les propriétés 1 et 2 sont immédiates ; leur démonstration est laissée à titre d'exercice. 3 : On note $m_X = E[X]$ et $m_Y = E[Y]$. Alors,

$$\begin{aligned} Cov(X, Y) &= E[(X - m_X)(Y - m_Y)] \\ &= E[XY] - E[Xm_Y] - E[Ym_X] + m_X m_Y \\ &= E[XY] - m_X m_Y - m_Y m_X + m_X m_Y \\ &= E[XY] - E[X]E[Y]. \end{aligned}$$

4 :

$$\begin{aligned} Cov(aX+b,cY+d) &= E[(aX+b)(cY+d)] - E[aX+b]E[cY+d] \\ &= acE[XY] + adE[X] + bcE[Y] + bd \\ &\quad -acE[X]E[Y] - adE[X] - bcE[Y] - bd \\ &= ac(E[XY] - E[X]E[Y]) \\ &= acCov(X,Y) \end{aligned}$$

□

Exemple On reprend l'exemple 2 du paragraphe 5.2 : $f(x,y) = e^{-y}$, $0 < x < y$, et $f(x,y) = 0$ sinon. On calcule tout d'abord les espérances de X et de Y :

$$E[X] = \int_0^\infty \int_0^y xe^{-y} dxdy = \int_0^\infty e^{-y} \int_0^y x\, dxdy = \frac{1}{2}\int_0^\infty y^2 e^{-y}\, dy = 1,$$

et

$$E[Y] = \int_0^\infty \int_0^y ye^{-y}\, dxdy = \int_0^\infty ye^{-y} \int_0^y dxdy = \int_0^\infty y^2 e^{-y}\, dy = 2.$$

On calcule enfin

$$\begin{aligned} E[XY] &= \int_0^\infty \int_0^y xye^{-y}\, dxdy = \int_0^\infty ye^{-y} \int_0^y x\, dxdy \\ &= \frac{1}{2}\int_0^\infty y^3 e^{-y}\, dy = \left[-\frac{1}{2}y^3 e^{-y}\right]_0^\infty - \frac{3}{2}\int_0^\infty -y^2 e^{-y}\, dy \\ &= \frac{3}{2}\int_0^\infty y^2 e^{-y}\, dy = \frac{3}{2} \times 2 = 3. \end{aligned}$$

Donc finalement, $Cov(X,Y) = 3 - 2.1 = 1$.
Enfin, on a le théorème suivant.

Théorème 7.3 *Si X et Y sont deux variables aléatoires* **indépendantes**, *alors*

$$Cov(X,Y) = 0.$$

Preuve : On démontre tout d'abord que si X et Y sont indépendantes, alors $E[XY] = E[X]E[Y]$. En effet (on donne la démonstration dans le cas continu ; la démonstration dans le cas discret est similaire) :

$$\begin{aligned} E[XY] &= \int\int xyf(x,y)\, dxdy = \int\int xyf_X(x)f_Y(y)\, dxdy \\ &= \int xf_X(x)\, dx \int yf_Y(y)\, dy = E[X]E[Y]. \end{aligned}$$

Donc, $Cov(X,Y) = E[XY] - E[X]E[Y] = 0$. □

Attention ! L'inverse n'est pas vrai ! Ce n'est pas parce que la covariance entre X et Y est nulle que les deux variables aléatoires sont indépendantes. Pour s'en persuader, on construit l'exemple

suivant : X est une variable aléatoire uniforme sur $]-1,1[$, et $Y = X^2$. Alors, $E[X] = 0$ et $E[XY] = E[X^3] = 1/2 \int_{-1}^{1} x^3\,dx = 0$ car x^3 est une fonction impaire. Donc $Cov(X,Y) = E[XY] - E[X]E[Y] = 0$. Pour autant, les variables aléatoires X et Y ne sont pas indépendantes (elles sont liées par la relation $Y = X^2$), mais le lien ne peut pas être décrit linéairement entre ces deux variables.

La covariance décrit comment deux variables s'écartent, en espérance, de leur espérance respective, mais exprimée en unité produit, ce qui rend l'interprétation difficile. Afin de rendre cette interprétation plus aisée, on introduit le coefficient de corrélation linéaire.

Définition 7.5 *Le coefficient de corrélation linéaire est la grandeur*

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

C'est une grandeur sans unités, insensible à des transformations linéaires sur les variables aléatoires et dont la valeur est comprise entre -1 et 1. Ce sont ces propriétés que nous allons démontrer maintenant.

Propriété 7.4 *Soient a, b, c, d, quatre réels quelconques. Alors,*

$$\rho(aX+b, cY+d) = \frac{a}{|a|}\frac{c}{|c|}\rho(X,Y).$$

Preuve : Comme $Cov(aX+b, cY+d) = acCov(X,Y)$, $Var(aX+b) = a^2Var(X)$ et $Var(cY+d) = c^2Var(Y)$, le résultat découle de l'application de la définition. □

Ainsi, seul le signe du coefficient de corrélation linéaire est sensible à une transformation linéaire des variables aléatoires, mais pas sa valeur absolue.

Corollaire : Soit $a \neq 0$ un réel. Alors

$$\rho(aX+b, X) = \frac{a}{|a|} \quad (= 1 \text{ si } a > 0,\ = -1 \text{ si } a < 0).$$

La preuve en est immédiate si on remplace $cY+d$ par X dans la propriété précédente. Ce résultat est très intéressant car il indique que deux variables aléatoires linéairement liées ont un coefficient de corrélation égal à l'unité si elles varient dans le même sens ($a > 0$) et égal à -1 si elles varient de façon opposée ($a < 0$).

Propriété 7.5 *Si X et Y sont deux variables aléatoires indépendantes, alors $\rho(X,Y) = 0$.*

La preuve est immédiate car dans ce cas, $Cov(X,Y) = 0$.

Théorème 7.4 *On a toujours*

$$-1 \leq \rho(X,Y) \leq 1.$$

Preuve : On démontre tout d'abord un résultat intermédiaire, la relation de Cauchy-Schwarz :

$$E[UV]^2 \leq E[U^2]E[V^2],$$

où U et V sont des variables aléatoires de variance finie. On considère la combinaison linéaire $\lambda U + V$. Alors

$$E[(\lambda U + V)^2] = \lambda^2 E[U^2] + 2\lambda E[UV] + E[V^2] \geq 0$$

Comme cette grandeur doit être positive pour toute valeur de λ, il faut que le discriminant de cette expression soit négatif, ce qui entraîne que

$$4E[UV])^2 - 4E[U^2]E[V^2] < 0,$$

ce qui démontre la relation de Cauchy-Schwarz.
Maintenant, pour démontrer le théorème, on applique la relation à $U = X - E[X]$ et $V = Y - E[Y]$, ce qui entraîne que $Cov(X,Y)^2 \leq Var(X)Var(Y)$, donc que

$$|Cov(X,Y)|/\sqrt{Var(X)Var(Y)} \leq 1.$$

□

Il est temps de résumer et de synthétiser tous ces résultats. Le coefficient de corrélation linéaire est une mesure du lien **linéaire** entre deux variables aléatoires. Il est :
– nul si les variables aléatoires sont indépendantes ;
– toujours compris entre -1 et 1 ;
– égal à -1 ou 1 si la relation entre les deux variables aléatoires est linéaire.

Exemple On reprend l'exemple habituel : $f(x,y) = e^{-y}$, $0 < x < y$, et $f(x,y) = 0$ sinon. Nous avions déjà calculé que $Cov(X,Y) = 1$. Afin de calculer ρ, il faut d'abord calculer les variances :

$$E[X^2] = \int_0^\infty e^{-y} \int_0^y x^2 \, dxdy = \int_0^\infty \frac{y^3}{3} e^{-y} \, dy = 2$$

et donc $Var(X) = 2 - E[X]^2 = 2 - 1 = 1$;

$$E[Y^2] = \int_0^\infty y^2 e^{-y} \int_0^x dxdy \int_0^\infty y^3 e^{-y} \, dy = 6$$

et donc $Var(Y) = 6 - E[Y]^2 = 6 - 4 = 2$.
Finalement,

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{1}{\sqrt{1.2}} = \frac{1}{\sqrt{2}} = 0,707.$$

7.5 Espérance et variance de la somme de variables aléatoires

Théorème 7.5 *Soit une famille de variables aléatoires, $X_1, \ldots, X_n$, d'espérances finies. Alors*

$$E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n].$$

Preuve : On donne une preuve dans le cas continu pour $n = 2$. Le résultat s'étend par récurrence pour tout n. La démonstration dans le cas discret est similaire ; elle est laissée au lecteur à titre d'exercice.

$$\begin{aligned} E[X_1 + X_2] &= \int\int (x_1 + x_2) f(x_1, x_2)\, dx_1 dx_2 \\ &= \int\int x_1 f(x_1, x_2)\, dx_1 dx_2 + \int\int x_2 f(x_1, x_2)\, dx_1 dx_2 \\ &= E[X_1] + E[X_2]. \end{aligned}$$

□

Un corollaire de ce résultat est la formule suivante :

$$E[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} a_i E[X_i].$$

L'espérance mathématique d'une combinaison linéaire est la combinaison linéaire des espérances mathématiques. En d'autres termes, l'espérance mathématique est un opérateur linéaire. Par exemple, $E[X - Y] = E[X] - E[Y]$.

Théorème 7.6 *Soit une famille de variables aléatoires, $X_1, \ldots, X_n$, de variances et de covariances finies. Alors,*

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Cov(X_i, X_j).$$

Preuve : On note $\mu_i = E[X_i]$. Par le théorème précédent, on a $E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mu_i$. Alors,

$$\begin{aligned} Var\left(\sum_{i=1}^{n} X_i\right) &= E\left[\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i\right)^2\right] \\ &= E\left[\left(\sum_{i=1}^{n} (X_i - \mu_i)^2\right)\right]. \end{aligned}$$

Or on a l'identité suivante, quels que soient les nombres $b_1, \ldots, b_n$.

$$\left(\sum_{i=1}^{n} b_i\right)^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} b_i b_j = \sum_{i=1}^{n} b_i^2 + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} b_i b_j.$$

En appliquant cette identité à $b_i = (X_i - \mu_i)$, puis en prenant l'espérance, on arrive à

$$\begin{aligned} Var\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} E[(X_i - \mu_i)^2] \\ &\quad + \sum_{i=1}^{n} \sum_{j=i+1}^{n} E[(X_i - \mu_i)(X_j - \mu_j)], \end{aligned}$$

et donc au résultat recherché. □

Théorème 7.7 *Soit une famille de variables aléatoires $X_1, \ldots, X_n$, de variances finies et* **indépendantes** *les unes des autres. Alors,*

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i).$$

La démonstration est immédiate, en appliquant le théorème précédent, et en tenant compte du fait que $Cov(X_i, X_j) = 0$ si X_i et X_j sont des variables aléatoires indépendantes.

Nous allons maintenant présenter deux applications du théorème 7.7, très importantes en statistiques.

Loi binomiale Soit $Y \sim \mathcal{B}(n, p)$ une variable aléatoire binomiale de paramètres n et p. Alors il existe une famille de variables aléatoires, $X_1, \ldots, X_n$, de n loi de Bernoulli, indépendantes, et toutes de paramètre p, telle que :

$$Y = X_1 + \cdots + X_n.$$

En appliquant les théorèmes 7.5 et 7.7, on calcule aisément l'espérance et la variance de Y :

$$\begin{aligned} E[Y] &= \textstyle\sum_{i=1}^{n} E[X_i] = & np, \\ Var(Y) &= \textstyle\sum_{i=1}^{n} Var(X_i) = & np(1-p). \end{aligned}$$

Au passage, remarquons que ce résultat nous donne de manière immédiate l'espérance et la variance d'une loi binomiale de paramètres (n, p).

Moyenne arithmétique Soit $X_1, \ldots, X_n$ une famille de variables aléatoires indépendantes, de même espérance, μ, et de même variance finie, σ^2. On note $\bar{X}$ la moyenne arithmétique[9] des X_i :

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

[9] Attention à cette notation ! Dans la littérature statistique, la moyenne arithmétique est presque toujours notée $\bar{X}$. Cette notation, que l'on retrouve également sur les machines à calculer, n'est pas à confondre avec l'opération de complémentation sur les ensembles.

En appliquant les théorèmes 7.5 et 7.7, on calcule l'espérance et la variance de $\bar{X}$:

$$\begin{aligned} E[\bar{X}] &= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu \\ &= \mu, \end{aligned}$$

et

$$\begin{aligned} Var(\bar{X}) &= \frac{1}{n^2} Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n\sigma^2 \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

On note le résultat très important, qui nous servira beaucoup par la suite : la variance de la moyenne arithmétique, $\bar{X}$, diminue lorsque n augmente.

8 Convergences

8.1 Inégalités

Théorème 8.1 (Inégalité de Markov) *Soit X une variable aléatoire positive et d'espérance finie. Alors*

$$P(X \geq a) \leq \frac{E[X]}{a}.$$

Preuve : Soit Y la variable aléatoire définie de la façon suivante :

$$Y = \begin{cases} 1 & \text{si } X \geq a \quad \Rightarrow X/a \geq 1, \\ 0 & \text{sinon} \quad \Rightarrow X/a \geq 0. \end{cases}$$

On a donc toujours $Y \leq X/a \Rightarrow E[Y] \leq E[X/a]$. Or Y est une variable aléatoire de Bernoulli de paramètre $P(X \geq a)$, et donc $E[Y] = P(X \geq a)$, ce qui mène au résultat. □

Cette inégalité permet de borner $P(X \geq a)$ quand on connaît l'espérance, pour toutes les densités admettant une espérance finie. Ce résultat étant très général, il ne faut pas s'attendre à ce que ce soit une majoration intéressante en pratique.

Exemple Soit X est une variable aléatoire uniforme sur $[0, 10]$. Alors $E[X] = 5$.
Comparons $P(X \geq a)$ avec la majoration issue de l'inégalité de Markov :

a	$P(X \geq a)$	$E[X]/a$
2	0.8	$5/2$
8	0.2	$5/8$
10	0	$1/2$

Il faut noter que la majoration à la première ligne n'est guère utile puisqu'elle est supérieure à 1 ! Les deux dernières lignes montrent que la majoration est assez peu précise. En fait, le principal intérêt de l'inégalité de Markov n'est pas de nous fournir une majoration, mais un outil pour construire d'autres théorèmes.

Théorème 8.2 (Inégalité de Chebychev) *Soit X une variable aléatoire d'espérance μ et de variance σ^2 finies. Alors, pour tout $t > 0$,*

$$P(|X - \mu| \geq t) \leq \sigma^2/t^2.$$

Preuve : On applique l'inégalité de Markov à la variable aléatoire positive $(X-\mu)^2$, ce qui nous donne :

$$\begin{aligned} P((X-\mu)^2 \geq t^2) &\leq E[(X-\mu)^2]/t^2 \\ \iff P(|X-\mu| \geq t) &\leq \sigma^2/t^2. \end{aligned}$$

□

Cette inégalité étant issue de l'inégalité de Markov, elle ne fournira pas une meilleure majoration que la précédente. Elle a l'avantage cependant de fournir une majoration de $P(|X-\mu| \geq t)$ pour toute variable aléatoire (positive ou non), dès lors que l'espérance et la variance sont finies et connues.

Exemple Soit X une gaussienne $\mathcal{N}(0,1)$. Dans les tables on trouve que $P(|X| \geq 2) \simeq 0.05$. L'inégalité de Chebychev donne $P(|X| \geq 2) \leq 1/4$.

8.2 Loi des grands nombres

Théorème 8.3 (Loi faible des grands nombres) *Soit $X_1, X_2, \ldots$ une famille de variables aléatoires indépendantes et identiquement distribuées (c'est-à-dire, toutes de même loi de probabilité), d'espérance finie μ et de variance finie σ^2. On note $\bar{X}_n$ la moyenne arithmétique calculée sure les n premières variables aléatoires : $\bar{X}_n = \sum_{i=1}^{n} X_i/n$. Alors, pour tout $\epsilon > 0$,*

$$P(|\bar{X}_n - \mu| \geq \epsilon) \to 0 \quad \textit{lorsque} \quad n \to \infty.$$

Lorsque n augmente, la moyenne arithmétique de $X_1, \ldots, X_n$ s'écarte de l'espérance mathématique de plus de ϵ avec une probabilité de plus en plus faible. On dit que la moyenne arithmétique converge *en probabilité* vers l'espérance mathématique, et on note :

$$\bar{X}_n \xrightarrow{p} \mu.$$

Preuve : La démonstration de ce théorème est une application immédiate de l'inégalité de Chebychev, appliquée à $\bar{X}_n$. Nous avions vu au paragraphe 7.5 que

$$E[\bar{X}_n] = \mu \quad \text{et} \quad Var(\bar{X}_n) = \sigma^2/n.$$

D'après l'inégalité de Chebychev :

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \;\to\; 0$$

lorsque $n \to \infty$. □.

8.3 Théorème Central Limite

Ce théorème est donné sans démonstration, car celle-ci utilise des notions qui sortent du cadre de ce cours.

Théorème 8.4 (Théorème Central Limite) *Soit $X_1, X_2, \ldots$ une famille de variables aléatoires comme dans le théorème 8.3 et U une variable aléatoire gaussienne $\mathcal{N}(0,1)$, indépendante des X_i. Alors,*

$$P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq u\right) \to P(U \leq u) \textit{ lorsque } \quad n \to \infty.$$

Ce théorème dit que $\bar{X}_n$, centré et réduit (c'est-à-dire, après que l'on ait soustrait l'espérance et divisé par l'écart-type) converge vers une variable aléatoire gaussienne standard, $\mathcal{N}(0,1)$. La convergence de ce théorème est une convergence en loi, ce que l'on note :

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

La notation $\xrightarrow{\mathcal{L}}$ signifie que, lorsque $n \to \infty$, la fonction de répartition de la variable aléatoire à gauche tend vers celle de la variable aléatoire à droite.
Ce théorème est un résultat très fort, d'importance capitale en statistique car il permet de justifier des méthodes statistiques (estimation, test, etc...) lorsque n tend vers l'infini (ce qu'on appelle le comportement asymptotique des méthodes statistiques). Nous allons illustrer le TCL par quelques convergences très utiles en probabilité.

Convergence de la loi binomiale vers la loi gaussienne

Soit X_n une famille de variables aléatoires binomiales de paramètres (n,p). Nous avons vu à l'exemple 1 du chapitre 6 qu'une variable aléatoire binomiale pouvait se décomposer en une somme de n variables aléatoires de Bernoulli indépendantes de paramètre p que l'on notera Y_i :

$$X_n = \sum_{i=0}^{n} Y_i.$$

Ainsi, on peut appliquer le Théorème Central Limite à $X_n/n = \bar{Y}_n$, avec $E[X_n/n] = E[Y_i] = p$ et $Var(X_n/n) = Var(Y_i) = p(1-p)$. Alors,

$$\frac{X_n/n - p}{\sqrt{p(1-p)/n}} = \frac{X_n - np}{\sqrt{np(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

La loi gaussienne est beaucoup plus facile à utiliser que la loi binomiale, d'où l'intérêt de ce résultat. En pratique, on utilise l'approximation gaussienne de la loi binomiale dès que les erreurs sont acceptables, ce qui se produit lorsque l'ensemble des conditions suivantes sont vérifiées : $n \geq 30$, $p > 0,1$ et $np(1-p) > 5$. La figure 8 illustre cette convergence. On y a représenté le système de probabilité d'une loi binomiale $\mathcal{B}(n,1/2)$ avec n augmentant de $n = 2$ à $n = 20$. Pour $n = 20$, la densité de probabilité d'une gaussienne $\mathcal{N}(10,5)$ ayant même espérance et même variance est superposée. On voit que l'accord est déjà très bon.

Convergence de la loi de Poisson vers la loi gaussienne

Soit Y_λ une famille de variables aléatoires de Poisson de paramètre λ. Avec un raisonnement analogue, on montre que

$$\frac{Y_\lambda - \lambda}{\sqrt{\lambda}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

En pratique, on utilise souvent l'approximation gaussienne de la loi de Poisson lorsque $\lambda \geq 15$. La convergence de la loi de Poisson vers la loi gaussienne est représentée à la figure 9, où le système de probabilité d'une variable aléatoire de Poisson $\mathcal{P}(\lambda)$ a été représentée pour $\lambda = 1, 2, 5, 10$. Dans ce dernier cas, la densité d'une gaussienne $\mathcal{N}(10, 10)$ ayant même espérance et même variance qu'une $\mathcal{P}(10)$ a été superposée. Ici encore, on voit un assez bon accord.

Exemple On considère $Y \sim \mathcal{B}(60, 1/3)$. On va approcher $P(|Y - 20| \geq 7)$ de deux façons différentes.

– *Inégalité de Chebychev :* on observe que Y peut s'écrire comme la somme de 60 variables aléatoires de Bernoulli, $X_1, \ldots, X_{60}$, de paramètre commun $p = 1/3$. On note $\bar{X}$ leur moyenne arithmétique. Alors,

$$P(|Y - 20| \geq 7) = P(|\bar{X} - 20/60| \geq 7/60).$$

En remarquant que $20/60 = 1/3$ est le paramètre p de la loi de Bernoulli, l'application de la loi des grands nombres donne

$$P(|\bar{X} - 20/60| \geq 7/60) < \frac{\sigma^2}{t^2} = \frac{1/3.2/3}{60.(7/60)^2} \simeq 0,27.$$

– *TCL :* Par l'application du TCL, on utilise l'approximation

$$\frac{\bar{X} - p}{\sqrt{p(1-p)/n}} = \frac{Y - np}{\sqrt{np(1-p)}} = \frac{Y - 20}{\sqrt{120/9}} \simeq \mathcal{N}(0, 1).$$

Ainsi,

$$P\left(\frac{|Y - 20|}{\sqrt{120/9}} \geq \frac{7}{\sqrt{120/9}}\right) \simeq P(\mathcal{N}(0, 1) \geq 1,92) \simeq 0,05$$

par lecture des tables de la loi gaussienne standard.

Dans cet exemple, on remarque que la majoration donnée par la loi des grands nombres est nettement plus élevée que la probabilité donnée par le TCL. La loi des grands nombres n'utilise que la variance de la variable aléatoire et est valable pour n'importe quelle distribution, ce qui ne permet pas une grande précision. Le TCL donne une distribution approchée pour Y, ce qui permet une précision beaucoup plus importante.

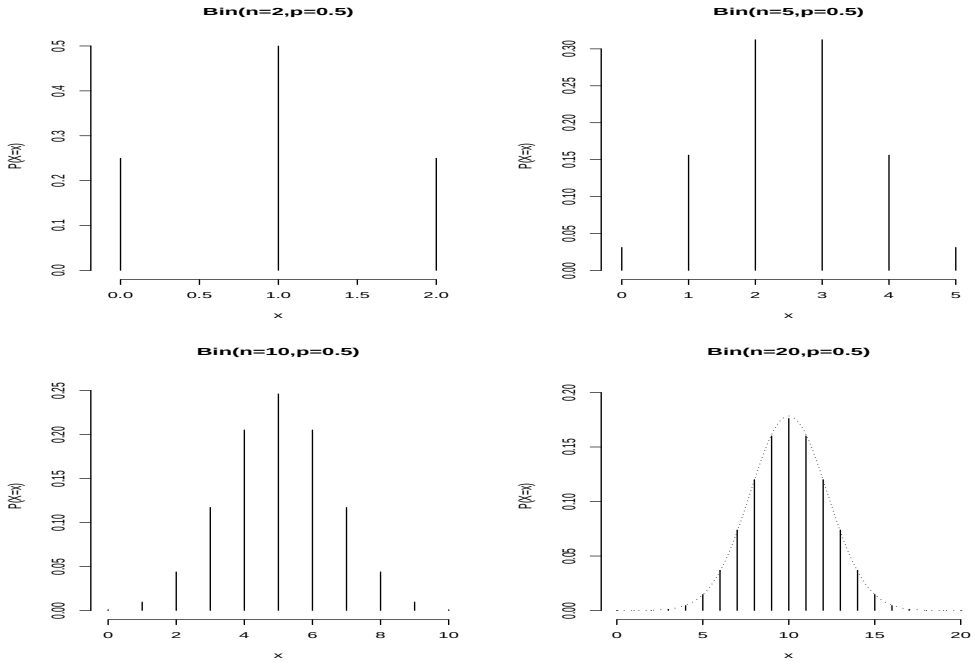

FIG. 8 – Convergence de la loi binomiale vers la loi normale.

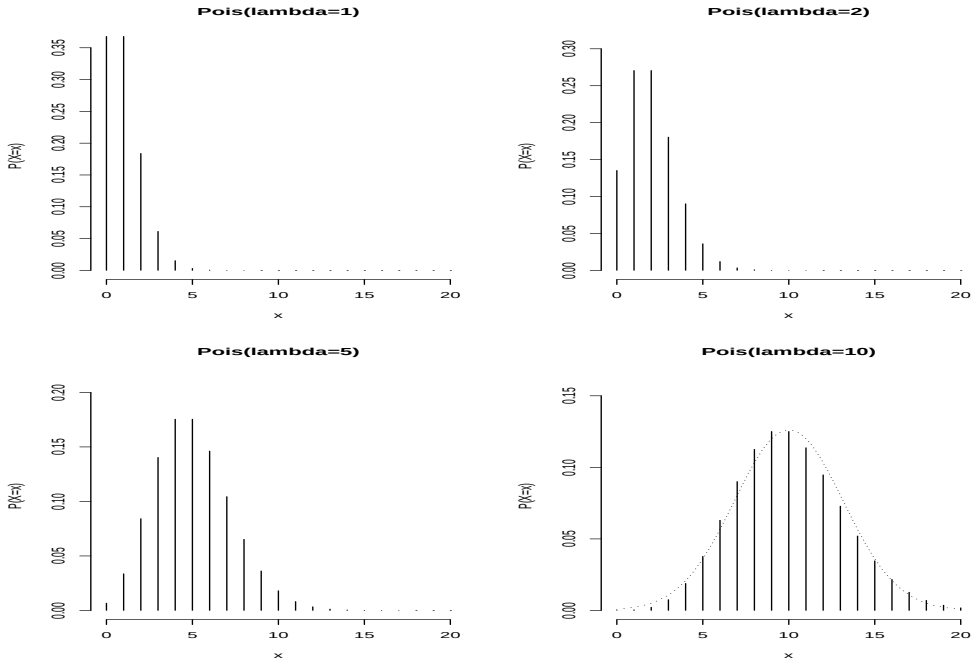

FIG. 9 – Convergence de la loi de Poisson vers la loi normale.

Conclusion

Ce dernier chapitre clôt la présentation rapide des probabilités. La loi des grands nombres permet de boucler la boucle de façon assez élégante. Nous avions vu au premier chapitre qu'avant que Kolmogorov ne propose une théorie mathématique basée sur les célèbres axiomes, on définissait auparavant la probabilité d'un évènement A comme la fréquence statististique d'apparition de cet évènement lorsque le nombre de répétitions de l'expérience aléatoire tend vers l'infini.

Nous venons de voir dans ce chapitre que l'axiomatique proposée par Kolmogorov (enrichie de la notion de variable aléatoire) permet de *démontrer* dans un cadre mathématique abstrait cette propriété empirique.

II. STATISTIQUES

Introduction

Les statistiques se définissent comme la collecte, la description et l'analyse de données. Ce qui est central en statistique, c'est le lien permanent entre les méthodes mathématiques et les données, lien qui apparaît dans chacune de ces trois tâches.

La question de la collecte des données est évidemment la plus fondamentale, car si la collecte est mal faite, et si les données sont de mauvaise qualité, alors l'analyse qui s'ensuivra sera forcément faussée. On raconte l'anecdote suivante. Vers la fin du XIX^{e}siècle (les statistiques n'en étaient qu'à leurs balbutiements), la ville de New-York voulut faire une enquête sur les petits commerces dans la ville. Plutôt que de faire une enquête exhaustive (un recensement), qui coûte très cher et qui prend beaucoup de temps, il fut décidé de procéder par sondage. Les enquêteurs décidèrent d'échantillonner les coins de rue (à New-York, cela revient à échantillonner sur une grille assez régulière) et aboutirent à une surestimation importante du nombre de commerces (pourquoi ?) et à des résultats complètement faux.

Les mini-sondages interactifs que l'on voit proliférer à la télévision ou sur Internet sont un exemple plus moderne de la même erreur. Il ne s'agit aucunement d'un sondage, car le résultat obtenu ne peut être extrapolé à la population entière : seul un sous-ensemble très particulier a souhaité participer, aux caractéristiques différentes de la population complète (les coins de rue dans un cas, les internautes visiteurs du site et qui prennent la peine de répondre dans l'autre). En termes statistiques, on dira que l'échantillon n'est pas représentatif de la population.

Toutefois, malgré l'importance de cette question, le problème de la collecte des données ne sera pas approfondi dans le cadre de ce cours.

Le premier chapitre de cette partie de cours est consacré à la statistique descriptive, qui regroupe l'ensemble des méthodes pour représenter et décrire des données (chapitre 9).
Les trois chapitres suivants sont ensuite consacrés à l'analyse des données, et trois aspects seront abordés :

- l'estimation des paramètres d'une population à partir d'un échantillon (c'est ce qu'on appelle les *statistiques inférentielles* (chapitre 10),
- les tests statistiques, qui mènent à *décider* du rejet ou non d'une hypothèse (chapitre 11),
- la modélisation des relations qui existent entre variables ; cette modélisation a le plus souvent un objectif prédictif : ce sera la régression (chapitre 12).

9 Statistiques descriptives

On peut définir les statistiques descriptives comme l'ensemble des méthodes qui permettent de synthétiser et de décrire de nombreuses données. La seule description objective des données est leur énumération. Si c'est possible lorsque le nombre de données est très petit, cela n'est rapidement plus possible lorsque le nombre de données est grand, voire simplement moyen : la simple énumération d'une trentaine de données est déjà très fastidieuse.

Synthétiser et résumer des données en un petit nombre de grandeurs est un exercice moins facile qu'il n'y paraît à première vue. Choisir une représentation, c'est renoncer à d'autres représentations possibles, et le choix de tel paramètre plutôt que tel autre n'est jamais neutre. Si la grandeur est toujours calculée selon des techniques parfaitement codifiées et reproductibles, le choix de cette grandeur entraîne une certaine subjectivité.
C'est la raison pour laquelle on dit souvent qu' « on peut faire dire ce qu'on veut aux statistiques ».
Nous verrons au paragraphe 9.6 un exemple qui illustre cet aphorisme.

9.1 Quelques définitions

Une **population** est l'ensemble des objets ou des personnes auxquels une étude statistique s'intéresse. Une population doit être définie par des critères ne laissant aucune équivoque : il faut être capable de dire sans erreurs possibles si tel objet ou telle personne appartient ou non à la population.

Exemples les élèves inscrits en première année de l'IUP d'Avignon durant l'année académique 2001–2002, les pièces usinées par l'atelier n de l'usine XYZ, du 2 février (06h00) au 5 février (19h00) 2001, les habitants d'Avignon redevables de la taxe d'habitation au 1erjanvier 2001 constituent une population au sens statistique.
Les habitants d'Avignon ou les personnes travaillant dans l'entreprise BIDULE, sans plus de précisions, ne constituent pas une population au sens statistique. En effet, qui doit-on compter, qui doit-on exclure ?
Un étudiant en cité universitaire rentrant toutes les fins de semaines dans sa famille, une stagiaire de 6 mois ou un SDF dormant dans les rues sont-ils des habitants d'Avignon ?
Un stagiaire percevant une compensation, un CDD de 1 mois ou un ingénieur consultant en régie sont-ils des personnes travaillant dans l'entreprise BIDULE ?
C'est parce qu'il est très difficile de répondre à ces questions[10] que « les habitants d'Avignon » sans plus de précisions ne constitue pas une population au sens statistique.

Une **unité statistique** — ou encore, un **individu** — est un élément de la population.

Un **échantillon** est un sous-ensemble de la population. Bien souvent il n'est pas possible de faire porter une enquête sur une population entière. On sélectionne alors un échantillon sur lequel on

[10] Sans compter que la réponse à ces questions peut être fluctuante selon les époques et/ou le contexte.

fait porter l'enquête. L'objectif des statistiques inférentielles est de tirer des conclusions sur la population à partir de l'étude de l'échantillon.

Les **caractères** décrivent l'unité statistique. Ainsi, le sexe, la catégorie socio-professionnelle (CSP), l'âge, le statut civil, le nombre d'enfants, le poids ou le revenu annuel sont des caractères pouvant décrire une personne. On distingue les caractères qualitatifs et les caractères quantitatifs.
Les **caractères qualitatifs** ne sont pas mesurables. Par exemple le sexe, la CSP ou le statut civil.
Les **caractères quantitatifs** sont mesurables. C'est le cas de l'âge, du nombre d'enfants, du poids et des revenus.
Parmi les caractères quantitatifs, on distingue les caractères discrets (qui peuvent s'énumérer : le nombre d'enfants, l'âge) et les caractères continus (qui ne peuvent pas s'énumérer : le revenu, le poids).
Les **modalités** sont les valeurs prises par un caractère qualitatif ou quantitatif discret. Le sexe n'a que deux modalités, les CSP de niveau 1 ont 5 modalités : agriculteur, ouvrier, employé, profession intermédiaire, cadres et professions intellectuelles supérieures (*sic*)[11].
La frontière entre discret et continu est assez floue. Ainsi l'âge peut parfaitement être considéré comme une variable continue. Le revenu est toujours un multiple du franc, et est donc une quantité intrinsèquement discrète. Cependant, elle est toujours considérée comme une variable continue. Ce qui fait souvent la différence, c'est la capacité à énumérer les modalités. Lorsqu'elles sont très nombreuses et rapprochées (typiquement, des francs, des grammes, etc...) on préfère une description avec des caractères continus.
Par analogie avec les variables aléatoires des probabilités, on utilise souvent le terme de **variable** à la place de caractère.

9.2 Description d'un caractère qualitatif ou quantitatif discret

Tableaux

Pour les caractères qualitatifs ou quantitatifs discrets, le tableau de contingence est une méthode de description objective et exhaustive car elle ne détruit aucune information. Soit p le nombre de modalités d'un caractère et n le nombre d'individus. On note $m_1, m_2, \ldots, m_p$ les modalités, $n_1, \ldots, n_p$ les effectifs dans ces différentes modalités et $f_1 = n_1/n, \ldots, f_p = n_p/n$ leurs fréquences statistiques. Le tableau 1 montre la structure d'un tableau de contingence.

Représentation en diagramme

Les diagrammes en bâtons ou en secteurs sont bien connus. Dans ces diagrammes, la hauteur des bâtons ou l'angle des secteurs sont proportionnels à la fréquence des modalités.

[11] Les CSP sont des catégories définies par l'INSEE ; les CSP de niveau 1 sont les moins détaillées ; les CSP de niveaux 2 et 3 ont beaucoup plus de modalités.

Modalité du caractère	effectif	fréquence
m_1	n_1	$f_1 = n_1/n$
$\vdots$	$\vdots$	$\vdots$
m_p	n_p	$f_1 = n_p/n$

TAB. 1 – Un tableau de contingence.

Ces diagrammes peuvent par exemple servir à illustrer le nombre de voitures immatriculées en France, où les différentes marques sont les modalités. Très souvent, on range les modalités aux effectifs très faibles dans une rubrique unique, intitulée « autres ».

Il existe une catégorie particulière de diagrammes, que l'on peut appeler pictogrammes. On y représente par exemple les parts de marchés des constructeurs automobiles par une voiture dont la taille est en relation avec la part de marché.
Ces pictogrammes sont très visuels et très accrocheurs. Cependant ils peuvent être très trompeurs, car on ne sait jamais si c'est la surface ou la hauteur du dessin qui est proportionnelle à la grandeur que l'on veut représenter.

9.3 Représentations des caractères continus

On considère que l'on a n individus dont on a mesuré les valeurs $x_1, \ldots, x_n$ d'un caractère continu. On va s'intéresser à deux représentations de ces n données.

Fonction de répartition empirique

La fonction de répartition empirique est l'équivalent de la fonction de répartition d'une variable aléatoire (cf. chapitre 4), mais ici elle est construite sur les valeurs $x_1, \ldots, x_n$. $\hat{F}(x)$ sera la proportion de valeurs inférieures ou égales à x :

$$\hat{F}(x) = \frac{1}{n} \# \left(\text{valeurs} \leq x \right).$$

La fonction de répartition empirique est une fonction constante par morceaux (fonction en escalier). Chaque saut correspond à une valeur x_i.

Nous allons illustrer cette notion à partir de données de pluviométries mesurées en Suisse le 8 mai 1986. Il s'agit de l'événement pluvieux qui a suivi l'accident de Tchernobyl, avec des nuages qui avaient été contaminés par les radionucléides. Nous avons 100 mesures de pluviométries mesurées aux stations météo du réseau Suisse. Ces données sont extraites d'une étude qui cherchait à évaluer la précision d'une estimation statistique de la pluviométrie en des points géographiques où elle n'avait pas été mesurée.

On voit à la figure 10 que les données vont de 0 à 600 1/10mm. On distingue des zones où la pente est plus forte, et des zones où elle est plus faible. Une pente élevée indique un grand nombre de données très voisines.

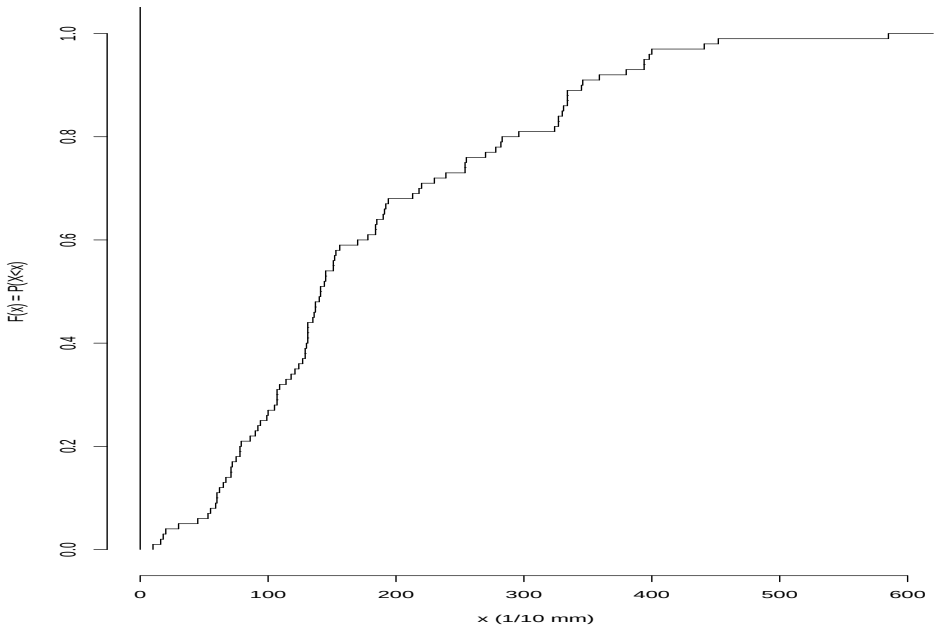

FIG. 10 – Fonction de répartition empirique de 100 données de pluie.

La fonction de répartition empirique est une représentation exacte des données, sans perte d'information, mais elle est peu parlante, et l'œil a du mal à interpréter ce genre de représentation. Aussi on préfère l'histogramme.

Histogramme

L'histogramme consiste à regrouper les n valeurs dans p classes (avec $p \leq n$), puis de traiter les classes comme des modalités que l'on représente par un diagramme en bâtons. Les classes sont des intervalles $[b_{i-1}, b_i[$, $i = 1, \ldots, p$. L'amplitude de la classe i est $a_i = b_i - b_{i-1}$ et son centre $c_i = (b_{i-1} + b_i)/2$. On résume alors les valeurs $x_1, \ldots, x_n$ dans un tableau statistique, identique au tableau 2.
L'habitude veut que les classes soient de même amplitude, mais cela n'est pas absolument indispensable. Il peut s'avérer utile, voire nécessaire de définir des amplitudes différentes. C'est par exemple le cas lorsqu'on étudie des revenus, car les classes des plus hauts revenus peut avoir une amplitude très grande.

Lors de la représentation graphique d'un histogramme, on trace des bâtons dont la hauteur est proportionnelle à la fréquence.

classe	centre	amplitude	effectif	fréquence
$[b_0, b_1[$	$c_1 = (b_0 + b_1)/2$	$b_1 - b_0$	n_1	$f_1 = n_1/n$
$[b_1, b_2[$	$c_2 = (b_1 + b_2)/2$	$b_2 - b_1$	n_2	$f_2 = n_2/n$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[b_{p-1}, b_p[$	$c_p = (b_{p-1} + b_p)/2$	$b_p - b_{p-1}$	n_p	$f_p = n_p/n$
Total	—	—	n	1

TAB. 2 – Tableau statistique lorsque les données sont groupées par classes.

L'histogramme est une représentation graphique très utilisée, mais il faut cependant faire attention au fait que sa forme peut être très sensible au choix du nombre et de la position des classes.

La figure 11 montre 6 histogrammes calculés sur les mêmes données de pluie. On voit très nettement une classe modale autour de 100 à 150 1/10mm. Augmenter le nombre de classes permet de visualiser plus finement la distribution des données. Ainsi, 7 classes ne permettent pas de distinguer une seconde classe modale qui apparaît autour de 300 1/10mm. Mais trop de classes entraînent un bruit tellement important qu'on ne visualise plus rien.

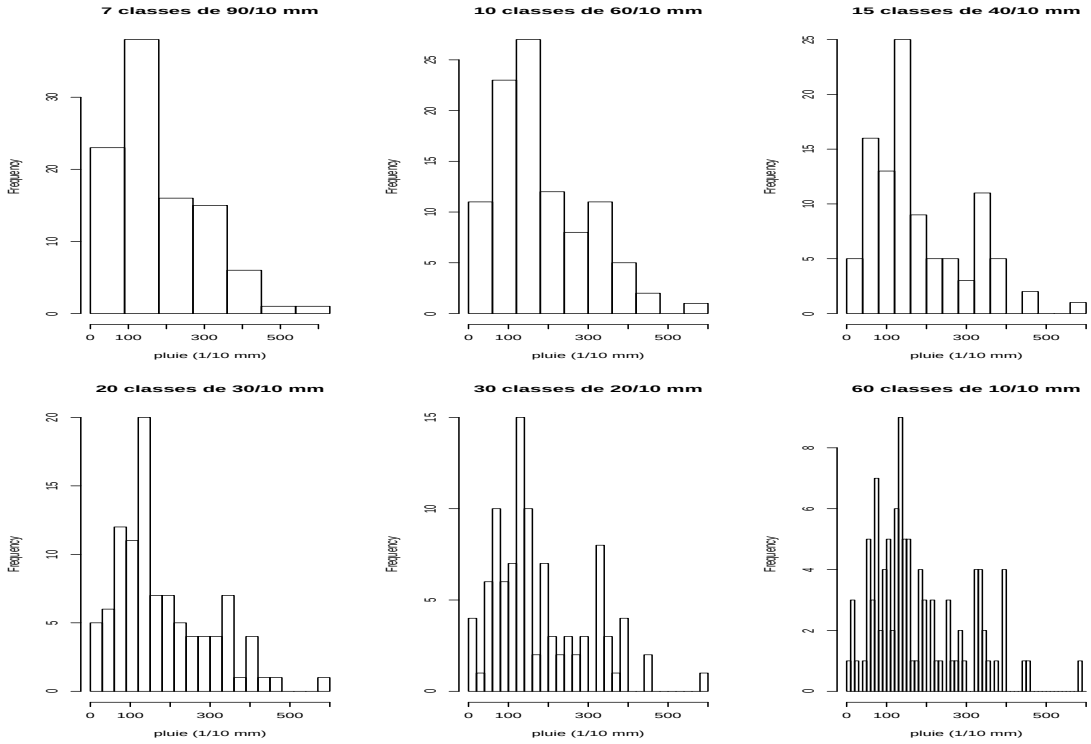

FIG. 11 – Histogrammes de 100 données de pluies

9.4 Caractéristiques de tendance centrale

La moyenne

C'est probablement le paramètre le plus connu des statistiques. On définit la moyenne arithmétique $\bar{x}$ (ou plus simplement, la moyenne) par :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Cette formule peut se décliner de plusieurs façons, selon les cas particuliers. Ainsi par exemple :

- Si on a un caractère discret à p modalités, la formule devient, en regroupant toutes les valeurs qui sont égales entre elles :
$$\bar{x} = \frac{1}{n} \sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} f_i x_i.$$
- Si on ne possède pas la liste des n valeurs, mais seulement un histogramme à p classes, on ne peut pas calculer une moyenne exacte, mais il est possible de calculer une approximation de la moyenne :
$$\tilde{x} = \frac{1}{n} \sum_{i=1}^{p} n_i c_i = \sum_{i=1}^{p} f_i c_i.$$
Cette approximation revient à considérer que toutes les données dans une classe ont la même valeur : le centre de cette classe.

La médiane

Par définition, la médiane est la valeur x_M qui partage les données en deux groupes d'effectifs égaux : le groupe des données supérieures à x_M, et le groupe des données inférieures à x_M. En d'autres termes, la médiane est la valeur x_M telle que

$$\hat{F}(x_M) = \frac{1}{2}.$$

On calcule x_M de la façon suivante : on ordonne les valeurs de façon croissante, et on note $x_{(1)} \leq \ldots \leq x_{(n)}$ les valeurs ainsi ordonnées.

- S'il y a un nombre de données n impair, alors la « donnée du milieu », c'est-à-dire
$$x_M = x_{((n+1)/2)}$$
est la médiane. En effet, dans ce cas, il y a bien $(n-1)/2$ données de chaque côté de la médiane.
- S'il y a un nombre pair de données, la médiane se trouve n'importe où entre les deux valeurs du milieu : $x_{(n/2)}$ et $x_{(n/2+1)}$. Par convention on prend souvent le milieu entre ces deux valeurs et
$$x_M = \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right).$$

Le mode

Le mode d'un caractère est sa modalité la plus fréquente. Le mode n'a réellement de sens que pour un caractère qualitatif ou quantitatif discret. Pour l'histogramme d'un caractère quantitatif continu, on parle parfois de classe modale pour la classe la plus fréquente. Toutefois, il faut se souvenir que la fréquence des classes dépend très fortement de leur définition (amplitudes et position des centres).

9.5 Caractéristique de dispersion

Après avoir défini quelques caractéristiques de tendance centrale, encore appelées caractéristiques de position, nous allons maintenant présenter quelques caractéristiques qui décrivent comment les valeurs sont réparties autour de la valeur de tendance centrale.

L'étendue

L'étendue, a, est simplement la différence entre la plus grande et la plus petite valeur :

$$a = x_{max} - x_{min} = x_{(n)} - x_{(1)}.$$

La variance

la variance, s^2, est définie comme la moyenne des écarts quadratiques à la moyenne :

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

La variance peut également se calculer de la façon suivante :

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.$$

En effet,

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{2}{n}\bar{x}\sum_{i=1}^{n} x_i + \bar{x}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}^2 + \bar{x}^2. \end{aligned}$$

L'écart-type, s, est la racine carrée de la variance. En statistique descriptive, il est préférable d'utiliser l'écart-type comme paramètre de description, plutôt que la variance, car celui-ci s'exprime dans les mêmes unités que la variable elle-même.

L'intervalle inter-quartile

On définit les quartiles x_{Q_1}, x_{Q_2} et x_{Q_3} à partir de la fonction de répartition empirique de la même façon que la médiane.

$$\hat{F}(x_{Q_1}) = \frac{1}{4}, \quad \hat{F}(x_{Q_2}) = \frac{1}{2}, \quad \hat{F}(x_{Q_3}) = \frac{3}{4}.$$

On note que $x_{Q_2} = x_M$. Le calcul de x_{Q_1} ou de x_{Q_3} se fait de la même manière que pour x_M.

L'intervalle-interquartile est :

$$IQ = x_{Q_3} - x_{Q_1}.$$

C'est l'intervalle qui contient 50% des données en en laissant 25% à gauche et 25% à droite.

L'intervalle inter-décile

On peut de même définir les déciles $x_{D_1}, \ldots, x_{D_9}$ qui partagent la fonction de répartition empirique en dix parties égales et l'intervalle inter-décile,

$$ID = x_{D_9} - x_{D_1},$$

qui contient 80% des données, en en laissant 10% à gauche et 10% à droite.

Coefficients de dispersion

Pour les variables positives, la dispersion est souvent rapporté à un paramètre de tendance centrale : $\sigma/\bar{x}$, IQ/x_M ou ID/x_M sont les coefficients de dispersion les plus souvents utilisés. Ce sont des grandeurs sans dimension.

9.6 Commentaires sur les caractéristiques

Quelles caractéristiques utiliser pour décrire et résumer des données ? La tentation est grande de les utiliser toutes, mais il faut prendre garde à ne pas générer plus d'informations que ne peut en assimiler le lecteur. D'un autre côté, choisir certaines caractéristiques plutôt que d'autres n'est pas toujours neutre, comme nous le montrerons dans l'exemple qui va suivre. L'équilibre n'est pas facile à trouver entre nécessité d'être bref et désir d'être complet.

Quelques remarques peuvent être faites :

- Il faut évidemment respecter des critères de cohérence dans l'utilisation de ces caractéristiques. Si on utilise la moyenne comme caractéristique de tendance centrale, il est normal d'utiliser la variance comme caractéristique de dispersion. Inversement, si on utilise la médiane comme caractéristique de tendance centrale, il faut utiliser les intervalles inter-quartiles ou inter-déciles comme caractéristique de dispersion.
- Moyenne et variance ont de bonnes propriétés mathématiques, mais sont sensibles à des valeurs extrêmes, mais rares. En termes statistiques, ont dit qu'elles sont peu robustes.

- Médiane et intervalle inter-quartiles sont au contraire très robustes (les valeurs maximales peuvent varier sans qu'elles n'en soient affectées). En revanche, elles ont de très mauvaises propriétés algébriques.
- En conséquence, on utilise souvent la médiane et l'intervalle inter-quartile ou inter-décile en statistique descriptive. Mais en statistique inférentielle, en test d'hypothèses, et en régression, on utilise plutôt moyennes et variances.

Exemple 1 Les données de pluie donnent les statistiques suivantes (en 1/10mm) :
- $\bar{x} = 180.2$ et $x_M = 141$;
- $x_{D_1} = 60$; $x_{Q_1} = 89$; $x_{Q_3} = 237.5$; $x_{D_9} = 345.5$.
- $s = 116.7$; $IQ = 148.5$; $ID = 285.5$.

Exemple 2 Une PME emploie 10 personnes, dont les salaires nets sont les suivants :

2	ouvriers	900 Euros,
4	ouvriers qualifiés	1100 Euros,
2	ingénieurs	1700 Euros,
2	directeurs	3300 Euros.

La moyenne des salaires vaut :

$$\bar{x} = (2 \times 900 + 4 \times 1100 + 2 \times 1700 + 2 \times 3300)/10 = 1620 \text{ Euros.}$$

La médiane vaut

$$x_M = \frac{1}{2}\left(x_{(5)} + x_{(6)}\right) = 1100 \text{ Euros.}$$

Compte tenu de ces statistiques, la direction dit : « dans notre entreprise, la moyenne des salaires est 1620 Euros et les salaires des dirigeants n'est que le double du salaire moyen ». Les syndicats disent : « dans cette entreprise, la moitié des salariés touchent 1100 Euros ou moins ; quant aux patrons, il empochent trois fois cette somme ».

Supposons maintenant que les directeurs décident de procéder à des augmentations. Les ouvriers et les ouvriers qualifiés sont augmentés de 100 Euros, les ingénieurs de 200 Euros, et les directeurs de 1000 Euros (on n'est jamais aussi bien servi que par soi-même).
Le nouveau tableau de salaire devient :

2	ouvriers	1000 Euros,
4	ouvriers qualifiés	1200 Euros,
2	ingénieurs	1900 Euros,
2	directeurs	4300 Euros.

La moyenne vaut maintenant $\bar{x} = 1920$ Euros et la médiane $x_M = 1200$ Euros. La direction peut clamer que le salaire moyen a augmenté de 300 Euros, mais les syndicats peuvent mettre en avant que seuls les directeurs ont une augmentation supérieure ou égale à cette moyenne, que

la médiane n'a augmenté que de 100 Euros et que 80% des salariés touchent moins que le salaire moyen.

Les deux points de vue étant statistiquement exacts, cet exemple a pour but de montrer que le choix d'une caractéristique pour résumer des données n'est pas neutre. Cela illustre également qu'en économie il peut se révéler peu judicieux d'utiliser des caractéristiques trop sensibles à des valeurs extrêmes, comme la moyenne.

10 L'estimation

10.1 L'estimation ponctuelle : généralités

En statistiques, on cherche souvent à estimer des paramètres ou des caractéristiques d'une population à l'aide d'un échantillon. Ainsi nous verrons par exemple que :
- la fréquence statistique de survenue d'un évènement est un estimateur de la probabilité de cet évènement ;
- la moyenne arithmétique calculée sur un échantillon est un bon estimateur de la moyenne arithmétique d'une population ;
- la variance calculée sur un échantillon est un estimateur de la variance d'une population.

Toutefois, il est parfois possible de définir plusieurs estimateurs pour une même caractéristique. Par exemple, si une densité de probabilité f est symétrique, la médiane x_M est aussi un estimateur de μ. Lequel doit-on choisir, x_M ou $\bar{x}$?

Il est donc nécessaire d'avoir une théorie de l'estimation pour répondre à cette question.

La cadre général est le suivant. On considère que l'on a un échantillon de n variables aléatoires idépendantes, $X_1, \ldots, X_n$, issu d'une loi de probabilité P_θ dont on cherche à estimer le paramètre inconnu θ. C'est le problème de l'**estimation ponctuelle**, dans le sens où on recherche **une valeur**, notée $\hat{\theta}$ (un point de $\mathbf{R}$) qui estime θ.

Définition 10.1 *On appelle* **estimateur** *une variable aléatoire $\hat{\theta}$ calculée à partir des X_i,*

$$\hat{\theta} = T(X_1, \ldots, X_n),$$

qui vise à estimer correctement θ et qui prend des valeurs dans un domaine acceptable pour θ.

Il faut remarquer qu'un estimateur est une fonction de variables aléatoires ; c'est donc une variable aléatoire, dont il est possible de calculer la densité, l'espérance et la variance si on connaît la loi P_θ et la forme de la fonction T.

Quelles sont les bonnes propriétés que doit avoir un estimateur $\hat{\theta} = T(X_1, \ldots, X_n)$?

Convergence

Il est indispensable que lorsque la taille de l'échantillon tend vers l'infini, $\hat{\theta}$ tende vers le paramètre θ :

$$\hat{\theta} = T(X_1, \ldots, X_n) \to \theta \text{ lorsque } n \to \infty.$$

Un estimateur non convergent sera toujours rejeté. Le fait qu'il converge (il est bon lorsque $n \to \infty$) ne nous dit cependant rien sur ce qui se passe lorsque n est fini (et on a rarement l'occasion d'aller jusquà $n \to \infty$).

Non biais

Lorsqu'on estime θ à l'aide de $\hat{\theta}$, on commet une erreur d'estimation, $\hat{\theta} - \theta$. En ajoutant et retranchant $E[\hat{\theta}]$, l'erreur d'estimation se réécrit :

$$\hat{\theta} - \theta = \left(\hat{\theta} - E[\hat{\theta}]\right) + \left(E[\hat{\theta}] - \theta\right).$$

Le premier terme est la fluctuation de $\hat{\theta}$ autour de son espérance. C'est une variable aléatoire liée au hasard de l'échantillonnage. Le second terme en revanche est une constante. C'est une erreur systématique que l'on commet en utilisant l'estimateur $\hat{\theta}$. C'est ce qu'on appelle le biais.

Définition 10.2 *Un estimateur $\hat{\theta}$ est un estimateur sans biais si :*

$$E[\hat{\theta}] - \theta = 0 \iff E[\hat{\theta}] = \theta.$$

Théorème 10.1 *Un estimateur sans biais est convergent ssi*

$$Var(\hat{\theta}) \to 0$$

quand $n \to \infty$.

Preuve : En effet, par application du théorème de Chebychev, $P(|\hat{\theta} - \theta| > \epsilon) \leq Var(\hat{\theta})/\epsilon$. Donc, si $n \to \infty$, alors $Var(\hat{\theta}) \to 0$ et $P(|\hat{\theta} - \theta| > \epsilon) \to 0$, c'est à dire $\hat{\theta} \xrightarrow{P} \theta$. □

Proposition 10.1 *Soit $X_1, \ldots, X_n$ un échantillon d'une population d'espérance μ. Alors la moyenne arithmétique $\bar{X}$ est un estimateur convergent et sans biais de μ.*

Preuve : En effet, nous avions vu au paragraphe 7.5 que

$$E[\bar{X}] = \mu \quad \text{et} \quad Var(\bar{X}) = \sigma^2/n.$$

La première égalité assure le non-biais. La seconde est équivalente à une loi des grands nombres. Comme $Var(\bar{X}) \to 0$ lorsque $n \to \infty$, on a bien que $\bar{X} \xrightarrow{p} \mu$. □

Proposition 10.2 *La fréquence statistique d'apparition, F, d'un évènement sur un échantillon est un estimateur sans biais de la propobabilité p de cet évènement dans la population.*

Pour montrer cette proposition, il suffit d'appliquer la proposition précédente aux variables aléatoires de Bernoulli d'apparition de cet évènement.

Proposition 10.3 *Soit $X_1, \ldots, X_n$ un échantillon d'une population de variance σ^2. La variance expérimentale*

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

est un estimateur biaisé de σ^2. L'estimateur sans biais est $\frac{n}{n-1}S^2$. Ces deux estimateurs sont convergents.

Preuve : Afin de montrer ce résultat, on a besoin du résultat intermédiaire suivant :

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2.$$

En effet :

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}(X_i - \mu - \bar{X} + \mu)^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + \sum_{i=1}^{n}(\bar{X} - \mu)^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - 2n(\bar{X} - \mu)^2 + n(\bar{X} - \mu)^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2.
\end{aligned}$$

Donc,

$$\begin{aligned}
E[S^2] &= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}E[(X_i - \mu)^2] - E[(\bar{X} - \mu)^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}Var(X_i) - Var(\bar{X}) \\
&= \frac{1}{n}.n\sigma^2 - \frac{\sigma^2}{n} \\
&= \frac{n-1}{n}\sigma^2.
\end{aligned}$$

On voit que $E[S^2] \neq \sigma^2$. Il y a un biais, qui vaut σ^2/n. Ce biais diminue avec n, et tend vers 0 lorsque $n \to \infty$. On dit que l'estimateur S^2 est asymptotiquement non biaisé. L'estimateur non biaisé de σ^2 est donc

$$\frac{n}{n-1}S^2.$$

La présence du biais vient de ce que, pour estimer la variance, on ne connaît pas la moyenne. Elle est seulement estimée à travers $\bar{X}$. Pour l'estimer, on «utilise» une fraction $1/n$ de chaque donnée X_i. Cette fraction déjà «utilisée» n'est plus disponible pour estimer la variabilité de la population, dont la variance est alors sous-estimée d'une fraction $1/n$.

Notons enfin que pour qu'un estimateur soit convergent, il faut nécessairement qu'il soit asymptotiquement sans biais.

Efficacité

Revenons à l'erreur d'estimation $\hat{\theta} - \theta$. Son espérance est le biais (qui peut être nul). Intéressons-nous maintenant à l'espérance du carré de l'erreur :

$$E[(\hat{\theta} - \theta)^2]$$

est l'erreur quadratique moyenne (EQM). Or :

$$\begin{aligned} E[(\hat{\theta} - \theta)^2] &= E[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)^2] \\ &= E[(\hat{\theta} - E[\hat{\theta}])]^2 + 2(E[\hat{\theta}] - \theta)E[(\hat{\theta} - E[\hat{\theta}])] + (E[\hat{\theta}] - \theta)^2 \\ &= E[(\hat{\theta} - E[\hat{\theta}])]^2 + (E[\hat{\theta}] - \theta)^2 \\ &= Var(\hat{\theta}) + (E[\hat{\theta}] - \theta)^2 \end{aligned}$$

car le second facteur du double produit est identiquement nul. On d'autres termes, on a l'équation suivante :

$$EQM(\hat{\theta}) = Var(\hat{\theta}) + Biais^2.$$

De deux estimateurs sans biais, le meilleur est celui dont la variance est la plus faible. Si l'estimateur est biaisé, il faut par contre tenir compte du terme de biais au carré, et il n'est plus évident que l'estimateur de plus faible variance soit celui dont l'EQM est le plus faible.

Définition 10.3 *Soient $\hat{\theta}_1$ et $\hat{\theta}_2$ deux estimateurs sans biais d'un paramètre θ. L'efficacité relative de $\hat{\theta}_1$ par rapport à $\hat{\theta}_2$ est le rapport $Var(\hat{\theta}_1)/Var(\hat{\theta}_2)$.*

Un estimateur $\hat{\theta}$ est une fonction de l'échantillon $X_1, \ldots, X_n$ et on ne peut calculer sa variance que si l'on s'est donné des hypothèses sur la distribution de la variable aléatoire X.

En général, la démarche est la suivante : on construit un modèle statistique, c'est-à-dire que l'on fait des hypothèses sur la loi des X (Poisson, exponentielle, gaussienne, etc...). Dans ce cadre, on peut alors calculer la variance de différents estimateurs $\hat{\theta}$ et choisir le plus efficace. On peut aussi se poser la question de savoir s'il existe un estimateur plus efficace que tous les autres. Cela revient à chercher l'estimateur sans biais de variance minimale.

La question est alors « comment fait-on pour trouver un estimateur ? Existe-t-il des guides pour en trouver ? Comment savoir si celui que j'ai trouvé est efficace (par rapport aux autres...) ». La réponse à ces questions fait l'objet des paragraphes suivants.

10.2 Méthode des moments

Le principe de la méthode des moments est très simple. Il consiste dans une première étape à calculer les moments théoriques (espérance mathématique, variance) issus de la loi de probabilité de la population. Ensuite, dans une seconde étape, on calcule les moments de l'échantillon (moyenne $\bar{X}$, variance S^2). Enfin, dans une troisième étape, on identifie ces deux modes de calcul afin d'en déduire l'estimation des paramètres.

Exemple 1 L'échantillon $X_1, \ldots, X_n$ provient d'une population dont la loi de probabilité est une loi géométrique de paramètre p. Il n'y a qu'un seul paramètre à estimer : on n'utilisera donc que le premier moment.

1. On a déjà vu que l'espérance mathématique d'une loi géométrique est $1/p$.
2. La moyenne arithmétique de l'échantillon est $\bar{X}$.
3. En identifiant moyenne arithmétique et espérance mathématique, on obtient un estimateur $\hat{p}$ de p :
$$\frac{1}{\hat{p}} = \bar{X} \iff \hat{p} = \frac{1}{\bar{X}}.$$

Exemple 2 L'échantillon $X_1, \ldots, X_n$ provient d'une loi gaussienne de paramètres (μ, σ^2). Ici, il y a deux paramètres à estimer ; on utilisera donc les deux premiers moments.

1. On a vu que l'espérance et la variance valent respectivement μ et σ^2.
2. La moyenne arithmétique et la variance empirique de l'échantillon sont respectivement $\bar{X}$ et S^2.
3. En identifiant, il vient de façon immédiate :
$$\hat{\mu} = \bar{X} \qquad \hat{\sigma}^2 = S^2.$$

Exemple 3 L'échantillon $X_1, \ldots, X_n$ provient d'un loi uniforme $[a, b]$. Ici, il y a également deux paramètres à estimer.

1. On a vu que l'espérance et la variance valent respectivement $(a+b)/2$ et $(a-b)^2/12$.
2. La moyenne arithmétique et la variance empirique de l'échantillon sont respectivement $\bar{X}$ et S^2.
3. En identifiant, on obtient les équations suivantes :
$$(\hat{a} + \hat{b})/2 = \bar{X} \quad \text{et} \quad (\hat{b} - \hat{a})^2/12 = S^2.$$
On obtient alors les estimateurs suivants :
$$\hat{a} = \bar{X} - S\sqrt{3} \quad \text{et} \quad \hat{b} = \bar{X} + S\sqrt{3}.$$

La méthode des moments est simple et intuitive. Toutefois, elle ne mène pas toujours aux estimateurs les plus efficaces.
Une autre méthode, issue d'une théorie plus complète permet de construire des estimateurs qui sont souvent optimaux.

10.3 Méthode du maximum de vraisemblance

Soit $X_1, \ldots, X_n$ une échantillon de loi f_θ, de paramètre inconnu θ. Afin d'alléger les notations, f_θ désigne une densité de probabilité ou la loi de probabilité d'une variable aléatoire discrète.

Définition 10.4 *On appelle* **vraisemblance** *la fonction*

$$L(X_1, \ldots, X_n; \theta) = f_\theta(X_1) \times \cdots \times f_\theta(X_n) = \prod_{i=1}^{n} f_\theta(X_i).$$

La vraisemblance est une fonction de θ quand les X_i sont fixés. Les valeurs de θ pour lesquelles les X_i ont une densité élevée sont des valeurs « vraisemblables » de θ pour l'échantillon considéré. Parmi toutes les valeurs possibles pour θ, il y en a une qui maximise la vraisemblance.

Définition 10.5 *On appelle* **estimateur du maximum de vraisemblance** *le valeur $\hat{\theta}$ qui maximise la fonction de vraisemblance.*

Cette démarche peut sembler à première vue curieuse. En fait, elle est parfaitement logique et correspond à une attitude courante de la vie de tous les jours. Face à une situation qui relève en partie du hasard (on attend une personne qui est en retard), on privilégie la cause qui a la plus grande probabilité (il y a eu des embouteillages) plutôt que les causes les moins vraisemblables (la personne a été enlevée par des extra-terrestres). Ici, on procède de même. On observe un échantillon et parmi les θ possibles (les causes possibles), on retient celui qui est le plus vraisemblable.

La technique la plus fréquente pour rechercher le maximum de vraisemblance consiste à rechercher la valeur $\hat{\theta}$ qui annule la dérivée première de la fonction de vraisemblance et de vérifier que la dérivée seconde est négative en ce point.

Remarque : bien souvent, il est plus facile de rechercher le maximum de la log-vraisemblance $\ln L(X_1, \ldots, X_n; \theta)$ que de la vraisemblance. Un petit calcul d'analyse montre facilement que le maximum d'une fonction et celui de son logarithme sont identiques. En effet $(\ln f(x))' = f'(x)/f(x)$. Et donc $f'(x) = 0 \iff (\ln f(x))' = 0$ si on suppose $0 < f(x) < \infty$.

Enfin, concernant l'estimateur du maximum de vraisemblance, on a le théorème suivant :

Théorème 10.2 *Soit $X_1, \ldots, X_n$ un échantillon d'une loi de probabilité de paramètre θ et $\hat{\theta}$ son estimateur du maximum de vraisemblance. Alors, sous certaines conditions assez générales l'estimateur du maximum de vraisemblance (sans biais) est convergent et de variance minimale. De plus, si on note*

$$I_n(\theta) = -E\left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right]$$

une quantité appelée information de Fisher*, alors,*

$$\sqrt{I_n(\theta)}(\hat{\theta} - \theta) \to \mathcal{N}(0, 1),$$

lorsque $n \to \infty$.

Exemple 1 **[Loi exponentielle]** Soit $X_1, \ldots, X_n$ un échantillon d'une loi exponentielle de paramètre λ. Alors la vraisemblance est :

$$\begin{aligned} L(X_1, \ldots, X_n; \lambda) &= \lambda^n \prod_{i=1}^{n} e^{-\lambda X} \\ &= \lambda^n \exp\{-\lambda \sum_{i=1}^{n} X_i\}. \end{aligned}$$

La log-vraisemblance est

$$\ln L(\cdots; \lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} X_i.$$

Pour maximiser la vraisemblance, on devra rechercher la valeur $\hat{\lambda}$ qui annule la dérivée première de la log-vraisemblance, et vérifier qu'en ce point la dérivée seconde est négative :

$$(\ln L(\cdots; \lambda))' = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i,$$

et donc

$$\frac{n}{\hat{\lambda}} - \sum_{i=1}^{n} X_i = 0 \iff 1/\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}.$$

La dérivée seconde

$$(\ln L(\cdots; \lambda))'' = -\frac{n}{\lambda^2}$$

est négative pour toute valeur de λ. $\hat{\lambda}$ est donc bien un maximum.
Dans cet exemple, on voit que l'estimateur de maximum de vraisemblance est la moyenne arithmétique de l'échantillon. Ici, la méthode des moments et celle du maximum de vraisemblance donnent le même résultat.

Lorsque la loi a p paramètres, la démarche reste identique à quelques nuances près. On recherche le maximum en annulant les p dérivées premières[12].

Exemple 2 **[Loi gaussienne]** Soit $X_1, \ldots, X_n$ un échantillon d'une loi gaussienne de paramètres (μ, σ^2). La vraisemblance est :

$$\begin{aligned} L(X_1, \ldots, X_n; \mu, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(X_i - \mu)^2}{2\sigma^2}\right\} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sigma^n} \exp\left\{-\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{2\sigma^2}\right\}. \end{aligned}$$

[12] Mathémtiquement, cette condition ne suffit pas. Il faut encore vérifier que la matrice des dérivées secondes est définie négative.

A nouveau, il est plus simple de travailler avec la log-vraisemblance :

$$\ln L(\cdots;\mu,\sigma) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{2\sigma^2}.$$

Les dérivées premières sont :

$$\begin{aligned}\frac{\partial \ln L(\cdots;\mu,\sigma)}{\partial\mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu),\\ \frac{\partial \ln L(\cdots;\mu,\sigma)}{\partial\sigma} &= -\frac{n}{\sigma}+\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^3}.\end{aligned}$$

Pour trouver l'estimateur de maximum de vraisemblance, on annule les dérivées premières, ce qui mène aux deux équations suivantes :

$$\begin{aligned}&\frac{\partial \ln L(\cdots;\hat\mu,\hat\sigma)}{\partial\mu} = 0 \quad \Rightarrow \sum_{i=1}^{n}(X_i-\hat\mu)=0 \quad \Rightarrow \hat\mu = \frac{1}{n}\sum_{i=1}^{n}X_i = \bar X,\\ &\frac{\partial \ln L(\cdots;\hat\mu,\hat\sigma)}{\partial\sigma} = 0 \quad \Rightarrow \sum_{i=1}^{n}(X_i-\hat\mu)^2 = n\hat\sigma^2 \quad \Rightarrow \hat\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\hat\mu)^2 = S^2.\end{aligned}$$

Dans cet exemple encore, les estimateurs du maximum de vraisemblance sont également les estimateurs donnés par la méthode des moments. Nous avions vu que S^2 est un estimateur biaisé de σ^2. Si on recherche un estimateur sans biais, on prendra plutôt $\hat\sigma = n/(n-1)S^2$. Le théorème 10.2 nous garantit que ces estimateurs sont les meilleurs possibles.

Exemple 3 **[Loi uniforme]** Dans ce dernier exemple, on considère que l'échantillon $X_1,\ldots,X_n$ provient d'une loi uniforme sur l'intervalle $[a,b]$. Nous savons que sur cet intervalle, la densité est constante et vaut $1/(b-a)$. En dehors de cet intervalle, elle est nulle. On introduit la fonction indicatrice $\mathbf{1}_A(x)$ de l'ensemble A qui vaut 1 si $x\in A$ et qui vaut 0 si $x\notin A$:

$$\mathbf{1}_A(x) = \begin{cases} 1 \text{ si } x\in A\\ 0 \text{ si } x\notin A.\end{cases}$$

La densité uniforme dans $[a,b]$ s'écrit :

$$f(x) = \frac{1}{b-a}\mathbf{1}_{[a,b]}(x).$$

Ayant introduit cette fonction indicatrice, on peut maintenant écrire la vraisemblance de l'échantillon $X_1,\ldots,X_n$:

$$\begin{aligned}L(X_1,\ldots,X_n;a,b) &= \prod_{i=1}^{n}\frac{1}{b-a}\mathbf{1}_{[a,b]}(X_i)\\ &= \left(\frac{1}{b-a}\right)^n\prod_{i=1}^{n}\mathbf{1}_{[a,b]}(X_i).\end{aligned}$$

Cette fonction n'est pas dérivable par rapport à a ou b. Il faut donc raisonner autrement. On voit que la vraisemblance sera maximale si $(b-a)$ est minimal. Pour cela, il faut b le plus petit possible et a le plus grand possible. Or b doit être plus grand que n'importe quel X_i et a plus petit que n'importe quel X_i (pourquoi ?). Il s'ensuit que :

$$\hat{a} = X_{min} \quad \text{et} \quad \hat{b} = X_{max},$$

en notant $X_{min} = \min(X_1, \ldots, X_n)$ et $X_{max} = \max(X_1, \ldots, X_n)$. On peut montrer que ces estimateurs sont légèrement biaisés :

$$E[\hat{a}] = a + (b-a)/(n+1) \quad \text{et} \quad E[\hat{b}] = b - (b-a)/(n+1),$$

mais asymptotiquement sans biais (le biais tend vers 0 lorsque $n \to \infty$). Les estimateurs du maximum de vraisemblance sans biais sont :

$$\begin{aligned} \tilde{a} &= X_{min} - \frac{X_{max} - X_{min}}{n+1}, \\ \tilde{b} &= X_{max} + \frac{X_{max} - X_{min}}{n+1}. \end{aligned}$$

Il faut remarquer que la méthode des moments avait donné des estimateurs très différents. Il est évidemment possible de calculer la variance de ces différents estimateurs. Le résultat est que l'estimateur par maximum de vraisemblance est beaucoup plus efficace que celui issu de la méthode des moments.

10.4 Estimation par intervalle de confiance

D'une certaine façon, l'estimation ponctuelle est peu satisfaisante. Nous avons vu qu'un estimateur est une variable aléatoire qui dépend de l'échantillon. Ainsi, si l'on change d'échantillon, on obtient une valeur différente de l'estimateur. Un exemple courant est l'enquête d'opinion. Il arrive fréquemment que plusieurs instituts de sondages se livrent à des enquêtes concernant la cote de confiance de tel ou tel personnage politique, et que l'on observe des différences entre les enquêtes. En dehors des effets liés à des différences de méthodologie, cela peut être simplement dû à des échantillons nécessairement différents, et ces différences peuvent ne refléter que la variabilité intrinsèque à l'échantillonnage aléatoire.

Une simple estimation ponctuelle n'apporte aucune indication sur une information pourtant essentielle : la précision de l'estimation. Ce n'est pas du tout la même chose de savoir que tel personnage politique qui se présente aux présidentielles est crédité de 53% d'intention de votes, plus ou moins 10% (rien n'est joué), ou plus ou moins 2% (ça semble gagné).

L'estimation par intervalle de confiance apporte cette information supplémentaire. Plutôt que d'estimer une valeur ponctuelle $\hat{\theta}$, on va encadrer θ par un intervalle $IC(\theta) = [\theta_1, \theta_2]$ qui contient, avec une forte probabilité, la valeur vraie mais inconnue θ.
Afin de déterminer cet intervalle, la procédure générale est la suivante :

1. On part tout d'abord d'un **niveau de confiance**, noté $1-\alpha$. Ce niveau de confiance correspond à la probabilité que l'intervalle contienne la vraie valeur :

$$P(\theta \in IC(\theta)) = 1-\alpha.$$

   En conséquence, α est la probabilité d'erreur que l'on s'accorde. Habituellement, on choisit $\alpha = 5\%$, mais selon le type de problème considéré, on peut choisir par exemple $\alpha = 1\%$.

2. On calcule une estimation ponctuelle $\hat{\theta}$ de θ, par maximum de vraisemblance par exemple. Bien souvent, par application du TCL, on connaît la loi de probabilité de $\hat{\theta}$ (ou une bonne approximation de celle-ci). Il est alors possible de rechercher les valeurs θ_1 et θ_2 de l'intervalle de confiance.

Cette démarche sera illustrée par le calcul de l'intervalle de confiance pour une proportion et pour une moyenne.

Intervalle de confiance d'une proportion

Le cadre est le suivant. Dans une population, une proportion p d'individus possède une certaine caractéristique (ils apprécient tel personnage politique, par exemple). On interroge un échantillon de n personnes, dont une proportion F possède la caractéristique en question. Si la population est très grande par rapport à la taille de l'échantillon, on peut la considérer infinie, et le modèle probabiliste qui correspond à cette situation est un échantillon de n variables aléatoires de Bernoulli de paramètre inconnu p.

Dans ce cadre, on a vu que $\hat{p} = F$ était un estimateur sans biais de p ; il est convergent. C'est l'estimateur des moments et du maximum de vraisemblance (exercice : montrer tout cela). On peut même montrer que c'est le plus efficace.

D'autre part, on avait également vu que lorsque $n \to \infty$, l'application du TCL donne

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

où $\mathcal{N}(0,1)$ est une variable aléatoire gaussienne standard, que l'on notera dorénavant Z.

On construit l'intervalle de confiance au niveau $(1-\alpha)$ de la façon suivante :

$$P(p_1 \leq p \leq p_2) = 1-\alpha \iff P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1-\alpha$$

et on recherche les valeurs p_1 et p_2 correspondantes.

Les valeurs $z_{\alpha/2}$ sont entièrement fixées par α. Elles figurent dans les tables statistiques habituelles, et sont en mémoire dans tous les logiciels de traitement statistique. Les valeurs les plus courantes sont résumées dans le tableau 3.

$1-\alpha$	$z_{\alpha/2}$
0,90	1,65
0,95	1,96
0,99	2,58

TAB. 3 – Quantiles usuels d'une variable aléatoire gaussienne standard.

On trouve les valeurs p_1 et p_2 par la suite d'équivalences suivantes :

$$-z_{\alpha/2} \leq Z \leq z_{\alpha/2}$$

$$\iff -z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2}$$

$$\iff \hat{p} - z_{\alpha/2}\sqrt{p(1-p)/n} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{p(1-p)/n}.$$

Comme on ne connaît pas la variance $p(1-p)$, on la remplace par $\hat{p}(1-\hat{p})$. Finalement, on aboutit à l'intervalle de confiance

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,,\, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

Exemple On interroge $n = 895$ personnes et 502 répondent positivement. Quel est l'intervalle de confiance au niveau $1-\alpha = 95\%$? L'application des formules donne : $\hat{p} = 502/895 = 0,561$.

$$p_1 = 0,561 - 1,96\sqrt{\frac{0,561.0,439}{895}} = 0,561 - 0,033 = 0,528$$

et

$$p_2 = 0,561 + 0,033 = 0,594.$$

Ainsi,

$$p \in [0,528; 0,594]$$

avec une probabilité 0,95.

Il n'est pas inutile de faire quelques commentaires au sujet de cette formule. L'amplitude de l'intervalle est donnée par le terme

$$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

– Ce terme augmente lorsque $1-\alpha$ augmente. C'est normal : à taille d'échantillon fixée, si l'on souhaite un plus grand niveau de confiance, il faut élargir l'intervalle. Notons qu'à la limite, l'intervalle qui correspond à un niveau de confiance *certain* est $p \in [0,1]$, ce qui n'est pas très informatif.
  Ainsi, il y a un équilibre à trouver entre précision et probabilité d'erreur.

- Ce terme diminue avec n, mais lentement (en $1/\sqrt{n}$). Pour diviser l'amplitude de l'intervalle par deux, il faut multiplier l'échantillon par 4 ! La taille de l'échantillon est directement liée au coût de l'enquête ; ici aussi, il faut arbitrer entre coût de l'enquête et coût de l'erreur d'estimation.
- Enfin, ce terme est fonction de $p(1-p)$. Ce terme de variance reflète l'incertitude liée au modèle. Il est maximal pour $p = 1/2$.

Dans le cas où p est proche de $1/2$ (sondage pour une présidentielle, par exemple), on peut faire l'approximation $\sqrt{p(1-p)} \simeq 1/2$ (pour $p = 0,4$, la vraie valeur est 0,4899). Par ailleurs, on arrondit très souvent 1,96 à 2. L'intervalle de confiance simplifié devient alors

$$[\hat{p} - 1/\sqrt{n}\,;\, \hat{p} + 1/\sqrt{n}].$$

Ainsi par exemple, si on veut une demi-amplitude égale à 0,03, il faut prendre un échantillon de taille $n = 1/0,03^2 = 1111$. On remarque que les sondages sont toujours réalisés avec un effectif d'environ 1000 personnes. Leur précision est donc de l'ordre de $\pm 3\%$. On note enfin que cette précision est atteinte *quelle que soit la taille de la population.* Ainsi, pour un niveau de précision donné, il faut interroger le même nombre de personnes aux États-Unis, en France ou au Lichtenstein.

Intervalle de confiance de la moyenne : variance connue

Nous avions vu que l'estimateur $\hat{\mu}$ de la moyenne μ d'une population est la moyenne arithmétique

$$\hat{\mu} = \bar{X}$$

des échantillons $X_1, \ldots, X_n$. On rappelle les résultats suivants :

$$E[\bar{X}] = \mu \qquad Var(\bar{X}) = \sigma^2/n.$$

Par ailleurs, en application du TCL,

$$\frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{L}} Z \sim \mathcal{N}(0,1).$$

On construit alors un intervalle de confiance d'un façon semblable au paragraphe précédent :

$$\begin{aligned} & -z_{\alpha/2} \leq Z \leq z_{\alpha/2} \\ \iff & -z_{\alpha/2} \leq \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \\ \iff & \hat{\mu} - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \hat{\mu} + z_{\alpha/2}\sigma/\sqrt{n}. \end{aligned}$$

Finalement, on aboutit à l'intervalle de confiance suivant :

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\,;\, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Intervalle de confiance de la moyenne : échantillon gaussien et variance inconnue

Si la variance est inconnue, on ne peut plus utiliser la formule précédente. Une autre formule est basée sur le résultat suivant, qui est donné sans démonstration :

Théorème 10.3 *Soit $X_1, \ldots, X_n$ un échantillon de loi gaussienne $\mathcal{N}(\mu, \sigma^2)$. Alors la quantité*

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

est une variable aléatoire distribuée comme une loi de Student à $(n-1)$ *degrés de liberté, noté* $T(n-1)$.

La densité d'une loi de Student à p degrés de liberté est donnée par la formule

$$f(t) = \frac{1}{\sqrt{n}B(1/2, p/2)(1 + t^2/p)^{(p+1)/2}}$$

ou $B(m, n)$ est la fonction Beta qui vaut

$$B(m, n) = \int_0^1 t^{m-1}(1-t)^{n-1}\, dt.$$

On peut montrer que $E[T(p)] = 0$ et $Var(T(p)) = p/(p-2)$. On a en outre le résultat de convergence suivant

$$T(p) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

lorsque $n \to \infty$. Les lois de Student sont tabulées ou préprogrammées, d'une façon tout à fait similaire à la loi normale. En pratique, on utilise la convergence de $T(p)$ vers une $\mathcal{N}(0,1)$ dès que $p > 50$.

Grâce à ce résultat, on construit un intervalle de confiance de la moyenne en suivant la même démarche qu'au cas précédent :

$$-t_{\alpha/2} \leq T(n-1) \leq t_{\alpha/2}$$

$$\iff -t_{\alpha/2} \leq \frac{\hat{\mu} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2}$$

$$\iff \hat{\mu} - t_{\alpha/2} S/\sqrt{n} \leq \mu \leq \hat{\mu} + t_{\alpha/2} S/\sqrt{n}.$$

Finalement, on aboutit à l'intervalle de confiance suivant :

$$\left[\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}\,;\, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right],$$

dont la formule est similaire à la précédente. Sa forme reste identique, mais il faut remplacer $z_{\alpha/2}$ par $t_{\alpha/2}$ et σ (inconnu ici) par S.

Il faut faire deux remarques :

1. L'hypothèse gaussienne est nécessaire pour que $\bar{X}$ soit gaussien et $\sqrt{n}(\bar{X} - \mu)/S$ une $T(n-1)$. Toutefois, si n est très grand, on peut s'affranchir de l'hypothèse gaussienne, et utiliser les propriétés du TCL. Dans ce cas on utilise les formules gaussiennes, car on profite de la convergence de $\bar{X}$ et S vers des $\mathcal{N}(0,1)$.
2. On peut vérifier dans les tables que $t_{\alpha/2} \geq z_{\alpha/2}$ pour tous degrés de liberté et toutes valeurs de α. Utiliser $t_{\alpha/2}$ au lieu de $z_{\alpha/2}$ revient donc à augmenter l'amplitude de l'intervalle de confiance. C'est le prix à payer en contrepartie de l'estimation de la variance, plutôt que de pouvoir travailler avec une variance connue (ce qui est tout de même un cas plus réaliste).

11 Les tests d'hypothèses

11.1 Présentation générale

En statistique décisionnelle, on est souvent amené à tester une hypothèse sur une population à partir d'un échantillon de cette population. On veut par exemple vérifier l'efficacité d'une nouvelle variété de céréales, ou celle d'un nouveau médicament. Entre le moment où est découverte une nouvelle molécule et sa mise sur le marché sous la forme d'un médicament, il se passe de nombreuses années (parfois jusqu'à dix ans), durant lesquelles on teste la molécule *in vitro*, puis sur des cobayes, puis sur des groupes humains, sains ou malades. La preuve de son efficacité doit être apportée, c'est-à-dire qu'il faut tester le médicament contre un placebo et contre d'autres médicaments anciens. Il doit évidemment être plus efficace que le placebo, mais aussi ne doit pas être moins efficace que les anciens médicaments.

Les tests sont effectués de la façon suivante : on forme deux groupes, l'un à qui l'on donne le placebo (ou les médicaments déjà existants) et l'autre à qui l'on donne le nouveau médicament. Afin d'éviter toutes influences psychologiques, ni les patients, ni les médecins ne savent ce que prennent les malades (c'est ce qu'on appelle les tests en double aveugle).

A la fin de l'essai, on compare les résultats mesurés sur le groupe ayant reçu le médicament et ceux du groupe témoin. On observe en général une différence. Mais comme chaque individu est différent, les résultats sur les deux groupes ont forcément une composante aléatoire (des essais sur des groupes différents auraient mené à des résultats différents). Afin de juger de l'efficacité du nouveau médicament, on attend une différence importante. En termes statistiques, on dit une différence *significative*. Une différence trop faible sera mise sur le compte du hasard, et le nouveau médicament n'aura pas fait ses preuves.

Formalisons maintenant un peu cet exemple. On souhaite tester une hypothèse (le nouveau médicament est plus efficace que l'ancien). En fait, comme la science est conservatrice par nécessité, on n'accepte une nouvelle hypothèse que si *on apporte la preuve que l'ancienne doit être rejetée*. Dans notre exemple, l'ancienne hypothèse est : le nouveau médicament n'est pas plus efficace que l'ancien. En statistique, l'ancienne hypothèse est appelée l'hypothèse neutre, ou l'hypothèse nulle[13], et notée H_0.

On va confronter cette hypothèse aux données de notre échantillon, et un peu comme dans un procès d'assises où il faut apporter la preuve de la culpabilité de l'accusé, il faudra apporter la preuve de la culpabilité de l'hypothèse nulle, c'est-à-dire la preuve que les données sont en contradiction avec H_0. Si les données sont trop en désaccord avec H_0, on sera amené à rejeter H_0. En revanche, si les données ne sont pas en désaccord avec H_0, on ne rejettera pas H_0.

[13] Non pas qu'elle soit nulle en soi, mais c'est une traduction malheureuse de l'anglais *null hypothesis* qui signifie hypothèse neutre.

Appliquée à l'exemple des médicaments, l'hypothèse nulle est « il n'y a pas de différence entre le nouveau et l'ancien médicament ». On compare les résultats obtenus dans les deux groupes. S'ils sont peu différents, on ne rejette pas H_0, donc on dit que le nouveau médicament n'apporte pas de gain significatif et on conserve l'ancien. S'ils sont très différents (en faveur du nouveau, bien sûr), on rejette H_0, c'est-à-dire qu'on abandonne l'ancien pour le nouveau.

En lisant attentivement l'exemple précédent, on aura remarqué que face à l'hypothèse nulle, qui représente le *statu quo*, il y a une autre hypothèse, celle qui nous intéresse vraiment, et qu'on aimerait voir valider par le rejet de H_0. Cette seconde hypothèse, notée H_1 est dite *l'hypothèse alternative*.

Un test statistique est une méthode mathématique et reproductible qui vise à prendre une décision quant au rejet ou non d'une hypothèse nulle H_0, face à une hypothèse alternative H_1.

La démarche générale pour élaborer un test d'hypothèse est la suivante :

1. Définir le cadre statistique (type d'échantillon, loi de probabilité etc,...)
2. Définir mathématiquement H_0 (par exemple, les espérances sont égales).
3. Définir mathématiquement H_1 (par exemple, les espérances sont différentes).
4. Fixer une valeur pour α.
5. En déduire le test et rejeter ou non H_0.
6. Ne pas oublier de conclure en rappelant la décision prise, avec son niveau de risque.

Appliquée à l'exemple du médicament, la démarche est donc la suivante :

1. Le nombre de personne guéries suit une loi binomiale de paramètres p (taux de guérison, inconnu) et$n$ (effectif du groupe sur lequel on réalise le test).
2. L'hypothèse H_0 est que le nouveau médicament a la même efficacité que les anciens, dont le taux de guérison est connu et vaut p_0. Donc,
$$H_0 \ : \ p = p_0.$$
3. L'hypothèse alternative est que le nouveau médicament est plus efficace que les anciens. Donc,
$$H_1 \ : \ p > p_0.$$
4. On choisit un niveau de risque, par exemple $\alpha = 5\%$.
5. Au niveau α on détermine une valeur p_c au delà de laquelle on ne considère plus que H_0 est vrai. Si la valeur observée, $\hat{p} > p_c$, on rejette H_0 ; sinon on ne rejette pas H_0.
6. On conclut sur l'efficacité du médicament.

Notons que dans la pratique des tests, on fait tous les calculs théoriques en supposant que H_0 est vrai. La raison en est double :

1. Tout d'abord, H_0 est notre hypothèse neutre. C'est dans le cadre de cette hypothèse que l'on vérifie si les données sont compatibles avec celle-ci.
2. En relation avec le point précédent, c'est souvent le seul cadre dans lequel on peut faire les calculs. En effet, dans l'exemple des médicaments, H_0 correspond à une efficacité identique et H_1 correspond à une efficacité supérieure. Il est facile de faire les calculs dans le premier cas (les moyennes sont égales), mais pas dans le second (la différence de moyenne n'est pas spécifiée).

Lorsqu'on prend une décision sur les hypothèses, on prend le risque de se tromper, car il se peut que, par hasard, notre échantillon indique une différence là ou il y en a pas, ou l'inverse. Deux types d'erreur sont possibles.

1. Rejeter H_0 alors que H_0 est vraie. C'est ce qu'on appelle le risque de première espèce, ou l'erreur de type I. Dans la théorie des tests statistiques, c'est le paramètre que l'on contrôle. On note :
$$\alpha = P(\text{commettre une erreur de type I}).$$
La probabilité α s'appelle le niveau de risque.
2. Ne pas rejeter H_0 alors qu'il aurait fallu. C'est l'erreur de seconde espèce, ou de type II (qui est plus difficile à évaluer). On note :
$$\beta = P(\text{commettre une erreur de type II}).$$
La quantité $1-\beta$ (la probabilité de ne pas commettre une erreur de type II) est appelée la *puissance* du test. Nous reverrons cette notion de puissance plus en détail au paragraphe 11.4.

Dans l'exemple choisi pour introduire la notion de test, on avait en tête que le nouveau médicament est meilleur que l'ancien (c'est l'hypothèse que l'on cherche à valider). On fait donc un test où les hypothèses sont :

$$H_0 : \text{efficacité équivalente},$$

contre :

$$H_1 : \text{meilleure efficacité pour le nouveau médicament}.$$

L'hypothèse H_1 ne considère pas que l'efficacité du nouveau médicament puisse être inférieure à l'ancien. L'hypothèse alternative est une inégalité simple. On dit que le test est **unilatéral** (il n'y a qu'un seul côté à l'inégalité).

On peut aussi considérer des tests pour lesquels l'hypothèse alternative est :

$$H_1 : \text{efficacité différente}.$$

Dans ce dernier cas, on ne spécifie pas dans quel sens a lieu l'inégalité : efficacité plus grande ou plus faible. On parle alors de test **bilatéral**.

Pour chaque type d'hypothèse à tester, il existe un test bien particulier à mettre en œuvre. Nous allons illustrer les tests statistiques à partir de deux cas : test de proportion et test de moyenne.

11.2 Test de proportion

Soit un échantillon de taille n d'une population dans laquelle la probabilité d'une certaine caractéristique est p (inconnue). Cette caractéristique est observée avec une fréquence statistique $\hat{p}$ dans l'échantillon. On veut tester :

$$H_0 \ : \ p = p_0,$$

contre

$$H_1 \ : \ p > p_0,$$

où p_0 est une valeur précise. Si $\hat{p}$ est « trop » différente de p_0, on rejette H_0. Dans le cas contraire, on ne rejette pas H_0. Il s'agit maintenant de construire le test statistique, c'est-à-dire de déterminer une valeur p_c, appelée valeur critique, telle que si $\hat{p} > p_c$, la différence $\hat{p} - p_0$ est jugée « trop » grande.

On a vu à de multiples reprises que $\hat{p}$ est l'estimateur optimal de p. On travaille sous H_0, ce qui signifie que $p = p_0$. Dans ce cas,

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

On détermine la valeur p_c à partir du taux d'erreur de première espèce que l'on accepte. Si on ne veut *jamais* commettre cette erreur, la seule solution consiste à ne *jamais rejeter* H_0, ce qui ne présente aucun intérêt. Donc, on commence par se fixer un niveau de risque α qui permet de déterminer p_c de la façon suivante :

$$\begin{aligned}
& P(\text{rejetter } H_0 \mid H_0 \text{ vraie}) && = \alpha \\
\iff \quad & P(\hat{p} > p_c \mid p = p_0) && = \alpha \\
\iff \quad & P\left(\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} > \frac{p_c-p_0}{\sqrt{p_0(1-p_0)/n}}\right) && = \alpha \\
\iff \quad & P\left(\mathcal{N}(0,1) > \frac{p_c-p_0}{\sqrt{p_0(1-p_0)/n}}\right) && = \alpha,
\end{aligned}$$

en appliquant le Théorème Central Limite. Donc,

$$\frac{p_c - p_0}{\sqrt{p_0(1-p_0)/n}} = z_\alpha$$

et

$$p_c = p_0 + z_\alpha \sqrt{p_0(1-p_0)/n},$$

où z_α désigne le quantile $1-\alpha$ de la loi gaussienne, c'est-à-dire,

$$P(\mathcal{N}(0,1) < z_\alpha) = 1-\alpha.$$

Donc, finalement, H_0 sera rejeté si :

$$\hat{p} > p_0 + z_\alpha \sqrt{p_0(1-p_0)/n}.$$

La valeur p_c est appelée la valeur critique du test, et l'intervalle $[p_c, +\infty[$, la zone critique.

Exemple Le taux de survie à un an d'une maladie grave vaut 35%. Un nouveau médicament est testé sur un groupe de 84 malades. Le suivi longitudinal de ces malades a montré que 38 malades étaient encore en vie après un an. Est-ce une amélioration significative du taux de survie ?

On va tester $H_0 \ : \ p = 35\%$ contre $H_1 : \ p > 35\%$. On observe 38 survies, donc $\hat{p} = 38/84 = 0,4524$ (un taux de 35% donne 29 survies). Le seuil critique est :

$$p_c = 0,35 + z_\alpha \sqrt{0,35 \times 0,65/84} = 0,35 + z_\alpha \times 0,0520.$$

- Si $\alpha = 0,05$, alors les tables statistiques donnent $z_\alpha = 1,645$ et $p_c = 0,35 + 0,0855 = 0,4355$. Comme $\hat{p} > p_c$, on rejette H_0.
- Si $\alpha = 0,01$ (ce serait par exemple le cas si le médicament a des effets secondaires importants ; on ne l'adoptera que si la différence de survie est très significative), alors $z_\alpha = 2,33$ et $p_c = 0,35 + 0,121 = 0,471$. Comme $\hat{p} < p_c$, on ne rejette plus H_0.

On voit donc que la décision statistique change en fonction du niveau de risque choisi. Comme il est évidemment hors de question de choisir le niveau de risque en fonction de la réponse souhaitée, la pratique statistique classique est de déterminer ce niveau *avant* de faire le test.

11.3 Tests de moyenne

Les tests de moyennes se font de façon différente selon que la variance est connue ou inconnue. Nous commencerons par le cas le plus simple, lorsque la variance est connue.

Variance connue

Soit $X_1, \ldots, X_n$ un échantillon de taille n de moyenne inconnue m et de variance connue σ^2. On souhaite faire le test bilatéral :

$$H_0 \ : \ m = m_0,$$

contre

$$H_1 \ : \ m \neq m_0,$$

où m_0 est une valeur précise.

La statistique de test sera la moyenne arithmétique $\bar{X}$, c'est à dire que si $\bar{X}$ est trop différent de m_0, on rejette H_0. Dans le cas contraire on ne rejette pas H_0.

Sous H_0, le TCL (voir paragraphe 8.3) donne la convergence suivante :

$$\frac{\bar{X} - m_0}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Au niveau de risque α, le test se construit par la suite d'équivalence suivante :

$$\begin{array}{rcll} & P(\text{ne pas rejetter } H_0 \mid H_0 \text{ vraie}) & = 1-\alpha \\ \iff & P(-z_{\alpha/2} < \frac{\bar{X}-m_0}{\sigma/\sqrt{n}} < z_{\alpha/2}) & = 1-\alpha \\ \iff & P(m_0 - z_{\alpha/2}\sigma/\sqrt{n} < \bar{X} < m_0 + z_{\alpha/2}\sigma/\sqrt{n}) & = 1-\alpha. \end{array}$$

Donc si $\bar{X}$ est dans l'intervalle

$$[m_0 - z_{\alpha/2}\sigma/\sqrt{n}\,,\, m_0 + z_{\alpha/2}\sigma/\sqrt{n}],$$

on ne rejette pas H_0. Dans le cas contraire, on rejette H_0.

Exemple Par définition, le Quotient Intellectuel (QI) d'une population a pour moyenne 100 et pour écart-type 20. On fait passer des tests à 36 étudiants de l'IUP d'Avignon et on observe une moyenne de 106,1. Au risque $\alpha = 5\%$, peut-on conclure que les étudiants de l'IUP d'Avignon sont différents du reste de la population ?

Telle que la question est posée, un test bilatéral s'impose. On va donc tester :

$$H_0 \ : \ m = m_0,$$

contre

$$H_1 \ : \ m \neq m_0.$$

Dans cet exemple, $m_0 = 100$, $\sigma = 20$, $n = 36$ et $\bar{X} = 106,1$. Au risque 5%, $z_{\alpha/2} = 1,96$. L'hypothèse $H_0 \ : \ m = 100$ n'est pas rejetée si $\bar{X}$ est compris dans l'intervalle

$$\begin{array}{rl} & [m_0 - z_{\alpha/2}\,\sigma/\sqrt{n}\,,\, m_0 + z_{\alpha/2}\,\sigma/\sqrt{n}] \\ = & [100 - 1,96.20/6\,,\, 100 + 1,96.20/6] \\ = & [100 - 6,5\,,\, 100 + 6,5] \\ = & [93,5\,,\, 106,5]. \end{array}$$

Comme $\bar{X} = 106,1$ est compris dans l'intervalle, on ne rejette pas l'hypothèse nulle au risque 5%.

Si on part du présupposé que les étudiants de l'IUP sont plus intelligents que la population générale, on peut faire le test unilatéral avec comme hypothèse alternative :

$$H_1 \ : \ m > m_0.$$

Dans ce cas, on ne rejette pas H_0 si

$$\bar{X} < m_0 + z_\alpha \, \sigma/\sqrt{n}.$$

Au risque 5%, on a $z_\alpha = 1,645$. Alors $m_0 + z_\alpha \, \sigma/\sqrt{n} = 100 + 1,645.20/6 = 105,5$. En pratiquant un test unilatéral on rejette H_0.

Cet exemple montre que le résultat d'un test peut dépendre du type de test que l'on fait : unilatéral ou bi-latéral. C'est un point sur lequel il faut toujours être vigilant.

Variance inconnue

Si la variance est inconnue et seulement estimée, on ne peut plus utiliser la convergence donnée par le TCL pour fair le test[14]. Dans ce cas, on doit remplacer l'utilisation de la loi gaussienne standard par une loi de Student à $(n-1)$ degrés de liberté.
Pour un test bilatéral par exemple, on ne rejette pas l'hypothèse H_0 si $\bar{X}$ est compris dans l'intervalle

$$[m_0 - t_{\alpha/2} S/\sqrt{n} \, , \, m_0 + t_{\alpha/2} S/\sqrt{n}]$$

où $t_{\alpha/2}$ est le quantile d'une loi de Student à $(n-1)$ degrés de liberté.

Exemple On reprend le même exemple que précédemment, mais ici, on suppose que l'écart-type $S = 20$ est calculé à partir des résultats sur les 36 étudiants. Au risque 5%, le test bilatéral nous donne l'intervalle de non-rejet suivant (avec $t_{0,025} = 2,03$ lu à partir des tables statistiques) :

$$[100 - 2,03.20/6 \, , \, 100 + 2,03.20/6] = [93,2 \, , \, 106,8].$$

Le test unilatéral au même niveau de risque, avec $t_{0,05} = 1,69$ donne l'intervalle de non-rejet :

$$]-\infty \, , \, 100 + 1,69.20/6] =]-\infty \, , \, 105,6].$$

A nouveau, on ne rejette pas H_0 pour un test bilatéral, mais on la rejette pour un test unilatéral.

11.4 Puissance d'un test

On rappelle que la puissance d'un test est la probabilité de ne pas commettre un erreur de type II, c'est-à-dire c'est la probabilité de rejeter H_0 lorsque H_1 est vraie. On voit tout de suite la difficulté qu'il y a à évaluer la puissance : il faut spécifier *numériquement* l'hypothèse alternative.

Nous allons illustrer la notion de puissance en reprenant le cas du test unilatéral d'une proportion. On rappelle qu'on a un échantillon de taille n d'une population présentant une certaine caractéristique avec une probabilité p. L'estimateur est la fréquence statistique de cette caractéristique dans l'échantillon, $\hat{p}$. On teste $H_0 \ : \ p = p_0$ contre $H_1 \ : \ p > p_0$. Pour faire le calcul

[14] sauf si l'échantillon est vraiment très important (ce qui est rarement le cas).

de la puissance, on doit se placer sous H_1. On suppose donc que la proportion réelle dans la population est $p > p_0$. On calcule la puissance :

$$\begin{aligned} 1-\beta &= P(\text{rejetter } H_0 \mid H_1 \text{ vraie}) \\ &= P\left(\hat{p} > p_c \mid p\right) \\ &= P\left(\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} > \frac{p_c - p}{\sqrt{p(1-p)/n}}\right) \\ &= P\left(\mathcal{N}(0,1) > \frac{p_c - p}{\sqrt{p(1-p)/n}}\right) \end{aligned}$$

en utilisant la convergence

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Or, $p_c = p_0 + z_\alpha\sqrt{p_0(1-p_0)/n}$. Donc :

$$\begin{aligned} 1-\beta &= P\left(\mathcal{N}(0,1) > z_\alpha\sqrt{\frac{p_0(1-p_0)}{p(1-p)}} - \frac{p-p_0}{\sqrt{p(1-p)/n}}\right) \\ &= 1-\Phi\left(z_\alpha\sqrt{\frac{p_0(1-p_0)}{p(1-p)}} - \frac{p-p_0}{\sqrt{p(1-p)/n}}\right) \end{aligned}$$

en notant $\Phi(\cdot)$ la fonction de répartition d'une variable aléatoire gaussienne standard. Dans cette expression, on voit que pour p_0 et pour α fixé, la puissance est une fonction de la proportion vraie, p. On représente en générale la puissance en fonction du paramètre de l'hypothèse alternative : c'est ce qu'on appelle la **courbe de puissance**. La figure 12 montre cette courbe pour l'exemple médical étudié au paragraphe 11.2. En trait plein est représenté la courbe de puissance pour $\alpha = 0,05$ et $n = 84$. On voit que pour $p = p_0$, la puissance est égale au niveau de risque α (cela se vérifie immédiatement en remplaçant p par p_0 dans l'expression de la puissance), puis la puissance augmente lorsque p s'éloigne de p_0. La puissance du test indique le pouvoir de discrimination du test, et la forme de la courbe nous indique que dans le cas où H_1 est vraie, on a d'autant plus de chances de rejeter H_0 que la vraie valeur p est éloigné de p_0.

En pointillé plein on a représenté la courbe de puissance lorsque le nombre de données double ($n = 168$). On voit que la puissance augmente avec le nombre de données. En pointillé léger figure la courbe de puissance lorsque le niveau de risque est $\alpha = 0,01$. Dans ce cas la puissance diminue, car si on cherche à diminuer la possibilité d'une erreur de type I, cela se fait nécessairement au détriment des erreurs de type II.

11.5 Seuil de significativité ou p-valeur

Jusqu'à présent, nous avons présenté les tests d'hypothèses comme une procédure visant à rejeter ou ne pas rejeter une hypothèse H_0 en présence d'une hypothèse alternative H_1, s'étant fixé *a priori* un niveau de risque α.

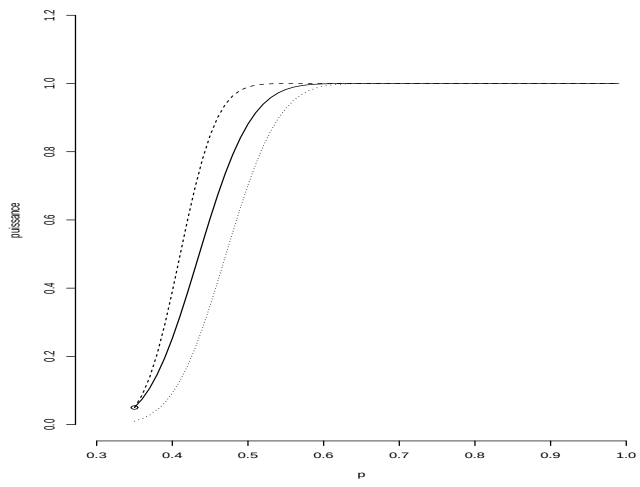

FIG. 12 – Courbes de puissance d'un test de proportion.

Une autre démarche est possible. Plutôt que de fixer α et décider de H_0, on calcule directement la probabilité de rejeter H_0 pour la valeur de la statistique observée. On appelle cette valeur un seuil de significativité, ou une p-valeur (ce dernier terme est la traduction de l'anglais *p-value*), et on la note α_0.

On reprend le cas étudié au paragraphe 11.2. Si on note p_{obs} la proportion observée, on a l'égalité suivante :

$$\alpha_0 = P(\text{rejeter } H_0 \mid H_0 \text{ vraie}) = P(\hat{p} > p_{obs} \mid p = p_0),$$

où on rappelle que $\hat{p}$ est la variable aléatoire représentant la fréquence statistique. Comme sous H_0 :

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

on a :

$$\begin{aligned} \alpha_0 &= P\left(\mathcal{N}(0,1) > \frac{p_{obs} - p_0}{\sqrt{p_0(1-p_0)/n}}\right) \\ &= 1 - \Phi\left(\frac{p_{obs} - p_0}{\sqrt{p_0(1-p_0)/n}}\right). \end{aligned}$$

Avec les valeurs de l'exemple médical, cela nous donne

$$\alpha_0 = 1 - \Phi\left(\frac{0,4524 - 0,35}{\sqrt{0,35.0,65/84}}\right) = 1 - \Phi(1,968) = 0,024.$$

Cette valeur est la probabilité d'avoir, sous H_0, un taux de survie supérieur à celui observé. Si on considère cette probabilité élevée, il ne faut pas rejeter H_0 et ne le rejeter que si on observe

un taux de survie encore supérieur. En revanche, si on considère cette probabilité comme faible, il ne faut pas rejeter H_0.

Rappelons-nous que H_0 est rejeté au niveau de risque 5%, mais n'est pas rejeté au niveau 1%. Il apparaît donc logique que la p-valeur, α_0, soit entre ces deux valeurs. En fait, α_0 est la valeur pour laquelle la décision s'inverse. Si $\alpha > \alpha_0$, on rejette H_0. Si $\alpha < \alpha_0$, on ne rejette pas H_0.

12 Modélisation bivariable

12.1 Description de données bivariables

On a un échantillon de n données $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Afin de visualiser le lien entre ces deux variables, l'outil de description le plus fréquemment utilisé est de représenter le nuage de points (x_i, y_i) sur un plan XY. On visualise alors immédiatement le lien entre les deux variables. Toutefois, cette façon de faire ne permet pas de quantifier ce lien. Le paramètre quantitatif le plus utilisé pour décrire ce lien est le coefficient de corrélation linéaire

$$r = \frac{1/n \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

où

$$s_x = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

est l'écart-type des valeurs marginales x_i et s_y est l'écart-type des y_i. On rappelle que pour des variables aléatoires (X, Y), le coéfficient de corrélation linéaire est définit par

$$\rho = \frac{E[(X - \bar{X})(Y - \bar{Y})]}{\sigma(X)\sigma(Y)}.$$

dont les propriétés avaient été présentée au paragraphe 7.4.
De façon similaire à ρ, r indique un lien linéaire entre les données (x_i) et (y_i). On a également $-1 \leq r \leq 1$. Si $|r| = 1$ il y a un lien linéaire exact entre les deux variables (d'où le nom). Plus r est proche de 1, plus on est proche d'un lien linéaire.
La figure 13 illustre quelques situations caractéristiques : en (a), les variables aléatoires sont indépendantes ; en (b), il existe un lien linéaire entre X et Y ; en (c), il y a un lien quadratique ; en (d), il y a un lien multiplicatif (qui est plus difficile à distinguer).

12.2 La régression linéaire

Supposons qu'après avoir dessiné le graphe des données (x_i, y_i), on observe une tendance linéaire et on souhaite ajuster une droite qui lie y à x. Plusieurs points de vue peuvent être adoptés pour résoudre ce problème, et il mènent (heureusement) tous au mêmes équations.

1. Les moindres carrés

On recherche a et b tels que

$$y_i^* = a + bx_i$$

soit « la meilleure » approximation de y_i. La quantité $y_i^* - y_i$ est l'erreur d'estimation commise lorsque y_i est estimé par y_i^*. On va chercher à minimiser l'erreur quadratique totale

$$Q = \sum_{i=1}^{n}(y_i^* - y_i)^2 = \sum_{i=1}^{n}(a + bx_i - y_i)^2.$$

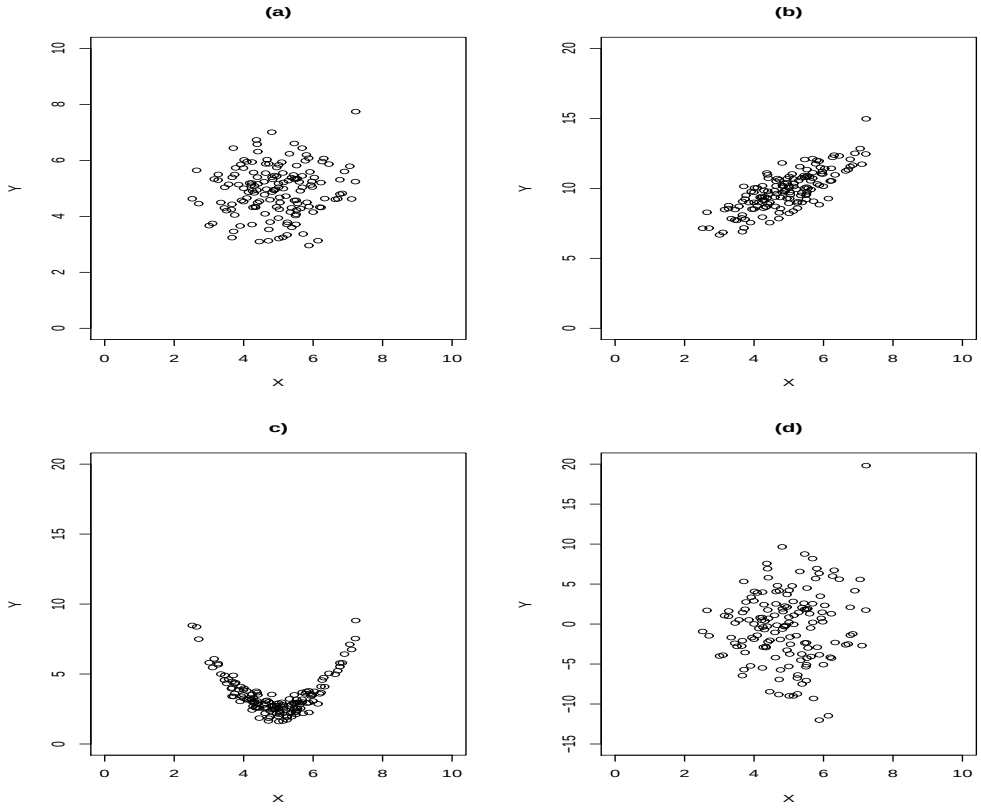

FIG. 13 – Représentation de couples de variables aléatoires (X, Y) : (a) indépendantes ; (b) lien linéaire ; (c) lien quadratique ; (d) lien multiplicatif.

Pour cela, on va rechercher les valeurs qui annulent les dérivées premières de Q par rapport à a et b. La dérivée par rapport à a donne :

$$\begin{aligned}\frac{\partial Q}{\partial a}(a,b)=0 &\iff 2\sum_i (a+bx_i-y_i)=0\\ &\iff a+b\bar{x}-\bar{y}=0\\ &\iff \bar{y}=a+b\bar{x},\end{aligned}$$

ce qui montre que la droite recherchée passe par le barycentre $(\bar{x},\bar{y})$ du nuage de points. La seconde dérivée donne :

$$\begin{aligned}\frac{\partial Q}{\partial b}(a,b)=0 &\iff 2\sum_i x_i(a+bx_i-y_i)=0\\ &\iff \sum_i x_i(\bar{y}-b\bar{x}+bx_i-y_i)=0\\ &\iff \sum_i x_i(\bar{y}-y_i)-b\sum_i x_i(\bar{x}-x_i)=0\\ &\iff r s_x s_y - b s_x^2 = 0\end{aligned}$$

qui s'annule lorsque

$$b = r\frac{s_y}{s_x}.$$

La droite recherchée, qui s'appelle la **droite de la régression de** y est finalement la droite

$$y=\bar{y}+r\frac{s_y}{s_x}(x-\bar{x}). \tag{2}$$

Si $r=0$, la pente est nulle, et $y=\bar{y}$ devient la meilleure estimation de y, indépendamment de x.

2. Point de vue probabiliste

Soit (X,Y) un couple de variables aléatoires. De la même façon qu'au paragraphe précédent, on va rechercher la droite $Y^*=a+bX$ qui donne la meilleure estimation possible de Y quand X est connue.
Le critère utilisé ici est de minimiser $Q=E[(Y^*-Y)^2]$. Comme au paragraphe précédent, on résout ce problème en annulant les dérivées premières de

$$\begin{aligned}Q=E[(Y^*-Y)^2] &= E[Y^{*2}-2YY^*+Y^2]=E[(a+bX)^2]-2E[Y(a+bX)]+E[Y^2]\\ &= a^2+2abE[X]+b^2E[X^2]-2aE[Y]-2bE[XY]+E[Y^2].\end{aligned}$$

Annuler les dérivées premières par rapport à a et b donne :

$$\begin{aligned}\frac{\partial Q}{\partial a}(a,b)=0 &\iff 2a+2bE[X]-2E[Y]=0\\ &\iff \mu_Y=a+b\mu_X.\end{aligned}$$

$$\begin{aligned}\frac{\partial Q}{\partial a}(a,b) = &\iff 2aE[X] + 2bE[X^2] - 2E[XY] = 0\\ &\iff 2aE[X] + 2b(Var(X) + \mu_X^2) - 2(Cov(X,Y) + \mu_X\mu_Y) = 0.\end{aligned}$$

En remplaçant dans cette dernière expression a par $\mu_Y - b\mu_X$, on arrive à

$$b = \frac{Cov(X,Y)}{Var(X)} = \rho\frac{\sigma_y}{\sigma_x}. \tag{3}$$

La première équation indique que Y^* est une estimation sans biais de Y, car on a $E[Y^*] = E[Y]$. Nous avions vu au paragraphe 10.1 qu'en absence de biais, minimiser l'écart quadratique ou minimiser la variance était équivalent. C'est le cas ici, et la seconde équation de la régression provient donc aussi de la minimisation de la variance $Var(Y^* - Y)$.

On voit donc que la régression linéaire d'une variable aléatoire (ici Y) par une autre (ici X) est obtenue en recherchant l'estimateur linéaire, sans biais, de variance minimale, ce qu'on appelle en anglais un BLUE (Best Linear Unbiased Estimator).

Il faut encore noter que l'équation 3 est formellement identique à 2. On voit donc, à travers ce calcul, le lien qui existe entre la minimisation d'une variance et le problème purement algébrique des moindres carrés.

3. Point de vue statistique

Enfin, il existe un troisième point de vue, statistique celui-ci. On considère le modèle statistique :

$$y_i = a + bx_i + \epsilon_i.$$

Dans ce modèle, on écrit que la donnée y_i est liée linéairement à x_i, à une erreur ϵ_i près. Ici, x_i et y_i ne sont pas aléatoires, c'est l'erreur qui est modélisée par une variable aléatoire, que l'on supposera gaussienne $\mathcal{N}(0,\sigma^2)$, c'est-à-dire d'espérance nulle et de variance σ^2. En outre, on supposera que les variables aléatoires ϵ_i et ϵ_j sont indépendantes, pour tous indices i et j. Les paramètres a et b sont inconnus, et on cherche à les estimer.

On va pour cela utiliser une méthode du maximum de vraisemblance. Comme

$$\epsilon_i = y_i - a - bx_i$$

est un échantillon de taille n d'une variable aléatoire gaussienne, la vraisemblance s'écrit :

$$L(\cdots; a, b) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y_i - a - bx_i)^2}{2\sigma^2}\right\}.$$

Le négatif de la log-vraisemblance est donc :

$$-\ln L(\cdots; a, b) = \frac{n}{2}\ln(2\pi) + n\ln\sigma + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2.$$

Maximiser la vraisemblance est équivalent à minimiser $-\ln L$, et il est immédiat que cela revient à minimiser $\sum_{i=1}^{n}(y_i - a - bx_i)^2$. En d'autres termes, l'estimation par maximum de vraisemblance est équivalente à la solution des moindres carrés. C'est entre autres raisons cette propriété mathématique qui fait que la densité gaussienne est la densité de loin la plus utilisée en statistiques.

On aboutit donc finalement aux estimateurs suivants :

$$\begin{aligned}
\hat{b} &= r\frac{s_Y}{s_X} \\
\hat{a} &= \bar{y} - \hat{b}\bar{x} \\
\hat{y}_i &= \hat{a} + \hat{b}x_i \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_i (y_i - (\hat{a} + \hat{b}x_i))^2
\end{aligned}$$

On peut énoncer le théorème suivant, sans en présenter la preuve qui sort du cadre de ce cours :

Théorème 12.1 *Soit le problème de régression linéaire présenté ci-dessus. Alors,*
– $\hat{a}$ et $\hat{b}$ sont les meilleurs (c.-à.-d. convergents, sans biais et de variances minimales) estimateurs de a et b.
– Par conséquent, $\hat{y}_i = \hat{a} + \hat{b}x_i$ est le meilleur estimateur de y_i.

Conclusions

Ce n'est là qu'un premier aperçu rapide et superficiel de la régression, qui occupe une place très importante en statistique. Nous ferons trois remarques rapides qui permettront peut-être de se faire une idée de la richesse de ce sujet.

- Le problème se généralise assez aisément au cas où on a plusieurs variables explicatives : $X^1, X^2, \ldots, X^p$ (ici, les exposants n'indiquent pas une puissance, mais un numéro de variables). Un problème qui fait toujours l'objet de recherches très actives est celui de la sélection des variables réellement explicatives lorsque celles-ci sont très nombreuses dans le jeu de données. Les études médicales sont des exemples typiques. On va par exemple mesurer sur des sujets la survie à une maladie en notant un très grand nombre de caractéristiques de ces sujets (variables médicales, habitudes alimentaires et d'hygiènes, variables biologiques, ...). On atteint rapidement plusieurs dizaines de variables, dont seulement certaines sont réellement pertinentes. Le but de l'analyse statistique sera de trouver lesquelles et d'estimer leur influence.
- Il est bien entendu possible de construire des tests visant à rejeter l'hypothèse nulle $H_0 : b = 0$ contre une alternative. Cela peut être fait dans le cas d'une seule variable ou lorsqu'on a plusieurs covariables.
- On n'est pas nécessairement limité aux relations linéaires. Ainsi des relations du type $y_i = ax_i^b$ ou $y_i = ae^{bx_i}$ se linéarisent facilement en passant au logarithme.

Par ailleurs, des méthodes appelées Modèle Linéaire Généralisé permettent de prendre en compte des relations plus complexes encore, tout en se ramenant *in fine* à des équations linéaires.