Chapitre 1

Probabilités

Dans le champs de l'observation, le hasard ne favorise que les esprits préparés.

(Louis Pasteur, 1854)

De la théorie des jeux à la modélisation financière, en passant par la biologie et bien sûr la physique (en particulier la mécanique quantique et la physique statistique), le hasard est omniprésent.

Par manque d'information, que l'objet étudié soit trop complexe ou que nos connaissances sur son état soient trop imprécises, il n'est pas toujours possible (ou même nécessaire) de prévoir le résultat d'une expérience avec certitude. C'est par exemple le cas d'une pièce lancée au jeu de pile ou face. Bien que le mouvement de la pièce soit déterminé par les lois de la mécanique, le résultat du jeu reste incertain. Mais derrière cette incertitude, se cache une quasi-certitude : en lançant 1000 fois la pièce, on sait que l'on obtiendra environ 500 fois "pile". Un événement *aléatoire* sera donc imprévisible, mais sa probabilité d'occurrence pourra être estimée en répétant l'expérience un grand nombre de fois.[1]

Bien que l'utilisation des probabilités soit quotidienne (météo, sondages, jeux, économie...), l'intuition est souvent trompeuse comme nous le verrons dans la suite de ce cours. Pour éviter les erreurs, un préalable est de définir une expérience idéale qui modélise au mieux la réalité. Par exemple, au jeu de pile ou face, on supposera par défaut que seuls deux résultats sont possibles (la pièce, *parfaite*, ne tombe pas sur sa tranche) et qu'ils ont la même probabilité $p = 1/2$ (n'ayant pas d'autres informations sur l'objet ou son propriétaire, tous les événements ont la même probabilité). Mais la pièce peut être (un peu) biaisée...

[1] Le mot *aléatoire* vient du latin *alea* et veut dire "jeu de dés", puis, par extension, hasard (qui vient de l'arabe *az-zahr* qui signifie aussi "dés"). Le hasard était déjà associé au jeu... En revanche, la notion d'ordre ou de régularité ressort de l'étymologie d'un synonyme du mot aléatoire : *stochastique*. En effet, ce mot plus savant venu du grec $\sigma\tau o\chi\alpha\sigma\tau\iota\kappa o\sigma$ n'a rien à voir avec le hasard, mais signifie plutôt "qui tend vers un but", puis "habile à conjecturer". Il est amusant de voir que le mot $\sigma\tau o\chi\alpha\sigma\tau\eta\sigma$ signifie le devin...

1.1 Événements, espace des observables et probabilités

1.1.1 Définitions

Depuis les travaux de A.N. Kolmogorov[2] (1933), les calculs de probabilités reposent sur la théorie de la mesure.[3] Dans la suite de ce cours, nous nous passerons de ce formalisme mathématique. Plus simplement, le calcul de probabilité est basé sur le trio suivant :

Événement

Un événement est le résultat possible d'une expérience idéale. En général, l'issue d'une expérience peut être décrit par un nombre entier (variable discrète, par exemple obtenue par un lancé de dé) ou par un nombre réel (variable continue, comme la taille d'un individu ou la vitesse d'une particule). Un événement peut être *élémentaire* (on dit aussi *simple*), par exemple faire 3 au dé, ou bien *composé* de plusieurs événements élémentaires (obtenir un nombre pair au dé, c'est-à-dire l'un de ces trois événements élémentaires : 2, 4 ou 6).

Espace des observables

L'*espace des observables* (également appelé *espace des événements*), noté Ω, est l'ensemble de tous les événements élémentaires d'une expérience.[4] Puisque l'espace des observables caractérise une expérience, il est fondamental de bien le définir, c'est-à-dire d'énumérer tous les résultats possibles. En revanche, la nature de ses points n'est pas importante : que l'on distribue des balles dans des boîtes ou des convives à des tables, le traitement est le même. Tout événement étant un ensemble d'événements élémentaires (au moins 1), on peut définir une structure algébrique munie de lois de compositions.

Probabilités

A tout événement élémentaire E_i de l'espace des observables Ω, on associe un nombre, la probabilité $p(E_i)$ d'obtenir l'événement E_i, telle que

- $p(E_i) \geq 0$ pour tout événement E_i,
- $Pr(E_i \text{ ou } E_j) = p(E_i) + p(E_j)$. La probabilité d'un événement composé est donc la somme des probabilités de tous les événements élémentaires le constituant,[5]
- $\sum_{E_i \in \Omega} p(E_i) = p(\Omega) = 1$ (l'événement certain est de probabilité 1).

[2] Andreï Nikolaevich Kolmogorov (1903-1987) mathématicien russe qui posa les bases de la théorie des probabilités sous sa forme axiomatique.

[3] On définit un espace de probabilités $(\Omega, \mathcal{C}, p)$, où Ω est un ensemble, $\mathcal{C}$ une *tribu* définie sur Ω, et p une mesure sur l'espace mesurable $(\Omega, \mathcal{C})$ telle que $p(\Omega) = 1$. Ces notions sont explicitées dans l'annexe A.1.

[4] En mécanique statistique, cet ensemble s'appelle l'espace des phases, un événement est alors la donnée de la position $\vec{q}_i$ et de la quantité de mouvement $\vec{p}_i$ de toutes les particules.

[5] On notera dans la suite $Pr(A)$, la probabilité de l'événement A décrit par l'argument.

On passe du vocabulaire probabiliste à celui de la théorie des ensembles à l'aide des lois de compositions suivantes. Soient A et B deux événements (a priori composés) :

- $A = \Omega$: l'événement A est certain et $p(A) = 1$ (obtenir un entier avec un dé).

- $A = \emptyset$: l'événement A est impossible et $p(A) = 0$ (obtenir 0 avec un dé).

- $A \subset B$: l'événement A implique l'événement B et $p(A) \leq p(B)$ (obtenir 2 implique un entier pair).

- L'événement $A^c = \Omega - A$ est l'événement complémentaire de A et $p(A^c) = 1 - p(A)$ (obtenir un entier pair et obtenir un entier impair).

- Loi de multiplication : $A \bigcap B$ (parfois notée $A.B$) signifie que les événements A *et* B sont réalisés.

  Les événements A et B sont dits *incompatibles* (ou *exclusifs ou disjoints*) si $A \bigcap B = \emptyset$ et donc $p(A \bigcap B) = 0$ (obtenir 2 et un nombre impair).

  Les événements compatibles A et B sont dits *indépendants* si et seulement si
  $$p(A \bigcap B) = p(A)p(B).$$
  (obtenir un nombre pair *et* obtenir un multiple de 3, c'est-à-dire 6, avec la probabilité $p = 1/2.1/3 = 1/6$).

- Loi d'addition : $A \bigcup B$ signifie que les événements A *ou* B sont réalisés ("ou" non exclusif, au moins l'un des deux) et
  $$p(A \bigcup B) = p(A) + p(B) - p(A \bigcap B).$$
  (obtenir un nombre pair *ou* obtenir un multiple de 3, c'est-à-dire 2, 4, 6 et 3 avec la probabilité $p = 1/2 + 1/3 - 1/6 = 2/3$).

  Cas particulier : si A et B sont *incompatibles* alors
  $$p(A \bigcup B) = p(A) + p(B).$$
  (obtenir 2 ou un nombre impair $p = 1/6 + 1/2 = 2/3$).

En pratique, on utilise une définition en fréquence : si on observe N événements *indépendants* lors d'une expérience dont n_i sont de type E_i, la probabilité de l'événement E_i est estimée par
$$p(E_i) = \lim_{N \to \infty} \frac{n_i}{N}.$$

1.1.2 Un exemple

Considérons l'ensemble des placements possibles de r balles discernables (on peut les distinguer les unes des autres) dans n boîtes. Chaque boîte peut recevoir un nombre illimité de balles. Explicitons les n^r événements élémentaires de Ω avec $r = n = 3$.

1.	$\{abc\|-\|-\}$	10.	$\{a\|bc\|-\}$	19.	$\{-\|a\|bc\}$
2.	$\{-\|abc\|-\}$	11.	$\{b\|ac\|-\}$	20.	$\{-\|b\|ac\}$
3.	$\{-\|-\|abc\}$	12.	$\{c\|ab\|-\}$	21.	$\{-\|c\|ab\}$
4.	$\{ab\|c\|-\}$	13.	$\{a\|-\|bc\}$	22.	$\{a\|b\|c\}$
5.	$\{ac\|b\|-\}$	14.	$\{b\|-\|ac\}$	23.	$\{a\|c\|b\}$
6.	$\{bc\|a\|-\}$	15.	$\{c\|-\|ab\}$	24.	$\{b\|a\|c\}$
7.	$\{ab\|-\|c\}$	16.	$\{-\|ab\|c\}$	25.	$\{b\|c\|a\}$
8.	$\{ac\|-\|b\}$	17.	$\{-\|ac\|b\}$	26.	$\{c\|a\|b\}$
9.	$\{bc\|-\|a\}$	18.	$\{-\|bc\|a\}$	27.	$\{c\|b\|a\}$

Table 1. *Les 27 placements de 3 balles numérotées a, b et c dans 3 boîtes.*

Par exemple, l'événement composé A, "un site est occupé plusieurs fois", est l'ensemble des événements 1 à 21. Et l'événement B, "le premier site n'est pas vide", est la réunion des événements 1, 4-15 et 22-27. On peut décider d'attribuer la même probabilité ($p_i = 1/27$) aux événements élémentaires[6] et s'intéresser par exemple à la probabilité de $A \bigcup B$ ou de $A \bigcap B$. Ainsi, $p(A) = 21/27$, $p(B) = 19/27$, $p(A \bigcap B) = 13/27$ et $p(A \bigcup B) = 1$. Les événements A et B ne sont donc pas indépendants.

1.1.3 Arrangements et dénombrements

Dans l'exemple précédent, le faible nombre de boîtes et de balles a permis de déterminer explicitement l'espace des observables. En général, on a recours à l'analyse combinatoire.

Soit un ensemble de n éléments, par exemple n boules numérotées dans une urne, on appelle *arrangement* de p éléments toute série *ordonnée* d'éléments *distincts* (tirage sans remise dans l'urne). Le nombre A_n^p d'arrangements de p éléments parmi n est :

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

En effet, les tirages sont indépendants et on a n possibilités pour le premier, $(n-1)$ pour le second et ainsi de suite. Un arrangement de n éléments parmi n s'appelle une *permutation* : $A_n^n = n! = n(n-1).(n-2)...2.1$ (par convention $0! = 1$).[7]

[6] Mais nous aurions pu choisir des balles "indiscernables", il n'y aurait alors plus que 10 éléments dans Ω. C'est la statistique quantique de Bose-Einstein.

[7] Si les éléments ne sont pas forcément distincts (tirage avec remise dans l'urne), le nombre de séries possibles est n^p, puisque l'on a n choix à chaque tirage indépendant.

Une collection *non ordonnée* de p éléments distincts parmi n est appelée *combinaison*. Le nombre C_n^p de combinaisons de p éléments parmi n est le coefficient binomial (également noté $\binom{n}{p}$) :

$$C_n^p = \frac{n!}{p!(n-p)!}. \tag{1.1}$$

Ainsi, le nombre de combinaisons possibles de p boules dans une urne qui en contient n (sans remise dans l'urne), indépendamment de l'ordre de sortie, est C_n^p. Prenons l'exemple d'un ensemble de trois éléments $\{a, b, c\}$. Les arrangements possibles de deux éléments sont : ab, ba, ac, ca, bc et cb. Soit $A_3^2 = 6$ arrangements. En revanche il y a $C_3^2 = 3$ combinaisons possibles : ab, ac et bc.

Quelques propriétés importantes :

- Symétrie : $C_n^{n-p} = C_n^p$
- Triangle de Pascal :

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

  On démontre cette relation en isolant un élément parmi les p. Il y a deux cas possibles : Soit cet élément appartient à la sélection, il reste alors à choisir $p-1$ éléments parmi $n-1$. Soit il n'appartient pas à la sélection et il faut choisir p éléments parmi $n-1$.

- Binôme de Newton : soient x et y deux réels,

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

  Cette formule se démontre par récurrence en utilisant le triangle de Pascal.[8]

[8] Pour $n = 0$ et $n = 1$ la formule est vérifiée. Par hypothèse de récurrence :

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

Exercice 1 *: les anniversaires*[†]

Quelle est la probabilité que les anniversaires de k personnes n'aient pas lieu le même jour ? En déduire la probabilité qu'au moins deux personnes aient le même jour d'anniversaire.

1.1.4 Probabilités conditionnelles

Axiome de Bayes :[9] Soient A et B deux événements compatibles de l'espace des observables Ω, on définit la *probabilité conditionnelle* $p(A|B)$ de réaliser l' événement A quand on a déjà réalisé l'événement B, par

$$p(A|B) = \frac{p(A \bigcap B)}{p(B)}. \tag{1.2}$$

C'est donc une probabilité définie sur le sous-ensemble $B \subset \Omega$. Ce qui donne la formule symétrique (et plus intuitive)

$$p(B|A)p(A) = p(A|B)p(B) = p(A \bigcap B).$$

Si les événements A et B sont *indépendants* alors $p(B|A) = p(B)$ (ou encore $p(A|B) = p(A)$) et on retrouve $p(A \bigcap B) = p(A)p(B)$.

Par exemple, la probabilité d'obtenir un multiple de 3 au dé (événement A : 3 ou 6) sachant que l'on a obtenu un nombre pair (événement B : 2, 4 ou 6) est $p(A|B) = p(A \bigcap B)/P(B) = (1/6)/(1/2) = 1/3$.

Généralisation : si les N événements A_i sont *exclusifs* et *exhaustifs* ($\Omega = \bigcup_{i=1}^{N} A_i$) et B est un événement quelconque alors

$$p(A_i|B) = \frac{p(B|A_i)p(A_i)}{\sum_{j=1}^{N} p(B|A_j)p(A_j)}.$$

Par exemple, A_1 est l'événement "la personne choisie est un homme", A_2, "la personne choisie est une femme" et B la personne choisie porte des lunettes.

[†]**Réponse** : Pour la première personne, il y a 365 possibilités, pour la seconde 364, pour la troisième 363...donc, puisque les événements sont indépendants :

$$p = \frac{364}{365}.\frac{363}{365}...\frac{365-(k-1)}{365} = A_{365}^{k}/365^{k}.$$

La probabilité pour qu'au moins deux personnes aient le même jour d'anniversaire est donc $1-p$. Pour k petit, on a :

$$1-p = 1-(1-\frac{1}{365})(1-\frac{2}{365})...(1-\frac{k-1}{365}) \simeq 1-(1-\frac{1+2+...+k-1}{365}) = k(k-1)/730.$$

Ainsi, pour $k = 20$ on trouve $1-p \simeq 0.5$, ce qui n'est pas si intuitif.

[9]Thomas Bayes (1702-1761), mathématicien et théologien anglais qui proposa une théorie des probabilités.

Exercice 2 *: les dès pipés*[†]

Soient 100 dés dont la moitié est pipée de telle sorte que la probabilité de tirer un 6 vaut $1/2$*. On choisit un dé au hasard, on le jette et on obtient un 6. Quelle est la probabilité que le dé choisi soit pipé ?*

1.2 Variable aléatoire et distribution de probabilité

Un événement aléatoire est un événement dont on ne peut pas déterminer l'issue, mais seulement lui associer une probabilité. Plus formellement, une *variable aléatoire* X est une fonction[10] définie sur l'espace des observables : à chaque point (événement) de cet espace, on associe un nombre, la probabilité que X vaille x, par une règle unique. Dans la suite, on notera en petites lettres les valeurs des variables aléatoires.

1.2.1 Distribution et densité de probabilité

Soit X une variable aléatoire qui peut prendre les valeurs discrètes $x_1, x_2, ...$, l'événement, éventuellement composé, tel que $X = x_j$ a une probabilité $P(x_j)$, où $P(x_i)$ est une fonction définie sur les valeurs de X appelée la *distribution de probabilité* de la variable aléatoire X.

Dans le cas d'une variable aléatoire continue, on définira la densité de probabilité[11] $P(x)$ par la probabilité que la variable aléatoire X ait une valeur comprise entre x et $x + dx$:

$$P(x)dx = Pr(x \leq X < x + dx).$$

La distribution/densité de probabilité vérifie les propriétés suivantes :

- $P(x_i) \geq 0$ pour tout x_i, ou $P(x) \geq 0$ pour tout x,
- $\sum_{x_i} P(x_i) = 1$, ou $\int P(x)dx = 1$ (normalisation).

On appelle *fonction de répartition* de la variable aléatoire X, la probabilité que X ait une valeur inférieure à x soit :

$$F_X(x) = Pr(X \leq x) = \int_{-\infty}^{x} P(y)dy.$$

La fonction $F_X(x)$ est donc non-décroissante et varie entre 0 et 1 telle que $F_X(-\infty) = 0$ et $F_X(+\infty) = 1$.

[†]**Réponse** : Soit A l'événement "choisir un dé pipé" ($p(A) = 1/2$) et soit B l'événement "tirer un 6" ($p(B) = 1/2.1/6 + 1/2.1/2 = 1/3$). La probabilité de tirer un 6 alors que le dé est pipé est bien sûr $p(B|A) = 1/2$ et la probabilité, si on a tiré un 6, que le dé soit pipé est $p(A|B) = p(B|A).p(A)/p(B) = 3/4$.

[10]Il vaudrait d'ailleurs mieux parler de fonction aléatoire plutôt que de variable aléatoire.

[11]Pour alléger les notations, on utilisera la même notation $P(x)$ pour une distribution et une densité de probabilité.

Considérons, par exemple, une pièce de monnaie lancée N fois. Le résultat étant "pile ou face", l'ensemble des observables est constitué de 2^N points représentant tous les résultats possibles lors des N lancés. Le nombre de lancés K qui ont donné un "pile" est une variable aléatoire qui peut a priori prendre les valeurs $0, 1, 2, ..., N$. La distribution de probabilité de la variable K est la loi binomiale (avec $p = 1/2$) que nous étudierons section 1.3.3. Autrement dit, la probabilité d'obtenir k fois un résultat "pile" sur les N lancés d'une pièce est $P_N(k) = C_N^k/2^N$.

1.2.2 Moments d'une distribution

Toute l'information sur une expérience dont l'issue est une variable aléatoire X est contenue dans la distribution de probabilité $P(x)$. Pour décrire sommairement cette fonction, on utilise sa *valeur moyenne* $\langle X \rangle$ (ou espérance mathématique $E(X)$) et sa *variance* $Var(X)$.

Valeur moyenne

La *valeur moyenne* $\langle X \rangle$ est définie par

$$\langle X \rangle = \sum_j x_j P(x_j) \quad \text{ou} \quad \langle X \rangle = \int x P(x)\, dx.$$

Par rapport à la connaissance de la distribution, la moyenne est une perte d'information : "Un homme s'est noyé en traversant une rivière dont la profondeur moyenne est de 30 cm"...

Pour décrire une distribution, on peut également introduire la valeur la plus probable (qui donne le maximum de $P(x)$) et la valeur médiane (telle que $F_X(x_{med}) = 1/2$).

Le terme "d'espérance mathématique" est très parlant dans l'exemple suivant : Au casino, un joueur mise 100 € sur l'un des 37 numéros de la roulette. Il a donc une chance sur 37 de gagner 35 fois sa mise et 36 "chances" sur 37 de la perdre. Son espérance de gain est donc $\langle X \rangle = \frac{1}{37}(3500) - \frac{36}{37}(100) \simeq -2.7$€...

Empiriquement, lorsqu'on réalise une expérience N fois, la valeur moyenne est estimée par

$$\langle X \rangle = \frac{1}{N} \sum_{p=1}^{N} x_p,$$

où x_p est la valeur de X obtenue la p^{eme} fois. Nous y reviendrons lorsque nous traiterons la loi des grands nombres au chapitre 1.4.

Soit X une variable aléatoire de densité de probabilité $P(x)$ et f une fonction, alors $F(X)$ est une nouvelle variable aléatoire qui peut prendre les valeurs $f(x)$. On appellera valeur moyenne de $F(X)$ la quantité

$$\langle F(X) \rangle = \int f(x) P(x)\, dx.$$

La densité de probabilité $\phi(f)$ de la variable aléatoire $F(X)$ est donnée par

$$\phi(f)df = P(x)dx,$$

soit,

$$\phi(f) = P[x(f)]\frac{1}{|f'(x)|},$$

si f est dérivable et monotone.[12]

Variance

Pour caractériser la largeur d'une distribution, ou l'écart à la moyenne,[13] on définit sa *variance* (ou écart quadratique moyen), notée $Var(X)$ ou σ^2, par la valeur moyenne du carré de l'écart par rapport à la moyenne :

$$Var(X) = \sigma^2 \equiv \langle (X - \langle X \rangle)^2 \rangle = \langle X^2 \rangle - \langle X \rangle^2 \geq 0.$$

La racine carrée de la variance, $\sigma = \sqrt{Var(X)}$, est appelée *écart type*. Par exemple, les températures moyennes annuelles à San Francisco et à Tokyo sont très proches (environ $13.5C^o$), mais l'écart type est de $1.4C^o$ dans la première et de $2.8C^o$ dans la seconde. La variance de la distribution d'un actif financier est appelée *volatilité*. Elle mesure l'instabilité de son cours.

Soient a et b deux réels, alors

$$\langle aX + b \rangle = a\langle X \rangle + b \quad \text{et} \quad Var(aX + b) = a^2 Var(X).$$

La translation ne modifie pas la variance d'une distribution.

On définit la variable *réduite* (normalisée) X^* associée à la variable aléatoire X par

$$X^* = (X - \langle X \rangle)/\sigma.$$

Ainsi $\langle X^* \rangle = 0$ et $Var(X^*) = 1$.

Moments

Plus généralement, une distribution $P(x)$ est caractérisée par son *moment d'ordre* n, où n est un entier positif ou nul, défini par :

$$\langle X^n \rangle = \sum_{x_j} x_j^n P(x_j) \quad \text{ou} \int x^n P(x)\, dx,$$

respectivement dans les cas de variable aléatoire discrète et continue. Le premier moment est donc la valeur moyenne et le second est lié à la variance. S'ils existent tous, leur connaissance permet en théorie de retrouver $P(x)$ (si P est analytique). La somme ou l'intégrale définissant le moment peut ne pas converger, dans ce cas, le moment n'existe pas. C'est le cas de la variance des lois larges que nous verrons plus loin. On remarquera que si le moment d'ordre n existe, alors tous les moments d'ordre inférieurs existent également.

[12] En effet, $\langle F(X) \rangle = \int \frac{f(x)}{|f'(x)|} P(x)df = \int f\phi(f)df$.

[13] On a bien sûr : $\langle X - \langle X \rangle \rangle = 0$.

1.2.3 Ensemble de variables aléatoires

On considère deux variables aléatoires continues X et Y dont les densités de probabilité sont respectivement $P(x)$ et $Q(y)$. La distribution de probabilité *jointe*, $H(x,y)$ (par unité de X et de Y) est définie par

$$H(x,y)dxdy = Pr\left(x < X < x+dx \text{ et } y < Y < y+dy\right).$$

Comme toute distribution de probabilité, $H(x,y)$ est normée. Si X et Y sont deux variables aléatoires indépendantes, alors :

$$H(x,y) = P(x)Q(y).$$

On définit la probabilité *conditionnelle* de l'événement $Y = y$ sachant que $X = x$ par

$$H(y|x) = \frac{H(x,y)}{\int H(x,y)\,dy}.$$

Remarquons que la fonction $H(y|x)$ ne dépend que de y pour un x donné, mais est différente de $Q(y)$.

La distribution de probabilité $P(x)$ est la distribution *marginale* de la variable X :

$$P(x) = \int_{y_{min}}^{y_{max}} H(x,y)dy.$$

Si X et Y sont deux variables aléatoires *indépendantes*, $S = X+Y$ est une nouvelle variable aléatoire dont la densité de probabilité est donnée par le produit de convolution[14]

$$H(s) = P * Q = \int P(t)Q(s-t)dt, \tag{1.3}$$

où $P(x)$ et $Q(y)$ sont respectivement les densités de probabilité de X et Y.

[14] À cause de l'indépendance des variables aléatoires X et Y, la probabilité que $X \in [x, x+dx]$ et $Y \in [y, y+dy]$ vaut $P(x)Q(y)\,dx\,dy$. La probabilité $H(s)\,ds$ est donc

$$H(s)\,ds = \int\int_{\mathcal{D}} P(x)Q(y)\,dxdy = \int_{-\infty}^{\infty} P(x)\,dx \int_{s-x}^{s+ds-x} Q(y)\,dy,$$

où le domaine $\mathcal{D}$ est défini par $\{x+y \in [s, s+ds]\}$. L'intégrale sur y se calcule par le théorème de la moyenne et vaut $Q(s-x)\,ds$.

Exercice 3 *: Dérivation de la loi de Maxwell* [15] *de distribution des vitesses.*[†]

On considère un gaz à l'équilibre. Soit $P(v_x, v_y, v_z)\, dv_x\, dv_y\, dv_z$ *la probabilité pour que dans l'espace des vitesses, l'extrémité du vecteur vitesse* $\vec{v}$ *soit dans le domaine* $[v_x, v_x + dv_x]$, $[v_y, v_y + dv_y]$, $[v_z, v_z + dv_z]$.

1) Déterminer $P(v_x, v_y, v_z)$ *à une constante près à l'aide des seules considérations de symétrie (homogénéité, isotropie, indépendance des 3 coordonnées). On introduira* $F(v_x)dv_x$ *la probabilité que* $v_x \in [v_x, v_x + dv_x]$.

2) Montrer que la probabilité pour que le module v *de la vitesse soit compris entre* v *et* $v + dv$ *est (distribution maxwellienne)*

$$W(v)dv = \frac{4}{\sqrt{\pi}} b^{3/2} v^2 e^{-bv^2} dv.$$

Calculer v_0*, la valeur la plus probable de* v*, sa valeur moyenne* $\langle v \rangle$*, puis sa valeur quadratique moyenne* $\langle v^2 \rangle$*. Dessiner* $W(v)$.

1.2.4 Corrélations et indépendance

Soient X et Y deux variables aléatoires continues de densité de probabilité $P(x)$ et $Q(y)$ respectivement. Posons

$$\Delta X = X - \langle X \rangle,$$

$$\Delta Y = Y - \langle Y \rangle.$$

On définit la *covariance* des variables X et Y par

$$cov(X, Y) = \langle \Delta X \Delta Y \rangle = \langle XY \rangle - \langle X \rangle \langle Y \rangle$$

et le *coefficient de corrélation* par

$$r(X, Y) = \frac{\langle \Delta X \Delta Y \rangle}{\sigma_X \sigma_Y} = \frac{cov(X, Y)}{\sigma_X \sigma_Y},$$

[15] James Clerk Maxwell (1831-1879) physicien anglais qui fonda la théorie de l'électromagnétisme et la théorie cinétique des gaz.

[†] **Réponse** :

1) Puisque le système est isotrope, les distributions des composantes des vitesses selon les trois directions de l'espace sont les mêmes. En supposant l'indépendance des probabilités (ce qui est critiquable !) on a :

$$P(v_x, v_y, v_z)\, dv_x\, dv_y\, dv_z = F(v_x)F(v_y)F(v_z)\, dv_x\, dv_y\, dv_z.$$

La distribution P ne peut dépendre que de la norme de la vitesse (isotropie) donc

$$P(v_x, v_y, v_z) = F(v_x)F(v_y)F(v_z) = f(v_x^2 + v_y^2 + v_z^2).$$

Il est facile de vérifier que seule l'exponentielle satisfait cette équation. On obtient donc

$$P(v_x, v_y, v_z) = Ae^{-b(v_x^2 + v_y^2 + v_z^2)},$$

où A et b sont > 0. La normalisation entraîne $A = (\frac{b}{\pi})^{3/2}$. Il ne reste plus qu'un paramètre, b. On montre en physique statistique que $b = m/2k_bT$, où T est la température, k_b la constante de Boltzmann et m la masse d'une molécule.

2) Le vecteur vitesse pointe alors dans le volume $4\pi v^2 dv$ décrit par la sphère de rayon v et d'épaisseur dv. On trouve alors à l'aide de l'annexe A.2 $v_0 = 1/\sqrt{b} = \sqrt{2k_bT/m}$, $\langle v \rangle = \sqrt{8k_bT/\pi m}$ et $\sqrt{\langle v^2 \rangle} = \sqrt{3k_bT/m}$.

où σ_X et σ_Y sont les écarts types de X et Y respectivement. Ainsi, si $Y = aX$, $cov(X,Y) = a\sigma_X^2 = \sigma_X\sigma_Y$, la corrélation est maximale et $r = 1$. En revanche, si les variables X et Y sont *indépendantes*, alors $r = cov(X,Y) = 0$ et donc

$$\langle XY\rangle = \langle X\rangle\langle Y\rangle.$$

Un exemple qui intéresse beaucoup l'industrie textile est la corrélation entre la taille d'un individu et la longueur de sa jambe (une étude statistique a montré que $r \simeq 0.8$).

En finance (et en physique), on s'intéresse en particulier aux corrélations temporelles. Supposons qu'une variable aléatoire soit donnée par une distribution de probabilité qui évolue dans le temps, $cov(X(t_1), X(t_2))$ donne une information sur les corrélations entre les instants t_1 et t_2.

Soient $X_1, ..., X_N$ des variables aléatoires de variances *finies* $\sigma_1^2, ...\sigma_N^2$, alors $S_N = \sum_{i=1}^N X_i$ est une variable aléatoire de variance

$$\boxed{Var(S_N) = \sum_{i=1}^{N} \sigma_i^2 + 2\sum_{i<j} cov(X_i, X_j)} \tag{1.4}$$

Cette relation permet de montrer que l'on a toujours[16] $|r| \leq 1$. Par ailleurs, la variance de la somme de variables aléatoires indépendantes est la somme de leurs variances.

L'indépendance de deux variables est donc une *condition suffisante, mais pas nécessaire,* pour qu'elles ne soient pas corrélées : l'absence de corrélation n'implique pas l'indépendance. Ainsi, des variables dépendantes peuvent ne pas être corrélées.

Exemple : Soit une distribution de probabilité $P(x)$ symétrique en $x = 0$ ($\langle X\rangle = 0$). Soit $Y = X^2$ une variable aléatoire, clairement dépendante de X ($\langle Y\rangle = \sigma^2$) alors $cov(X,Y) = \langle X^3\rangle - \langle X\rangle\sigma^2 = 0$ (car $\langle X^3\rangle = \langle X\rangle = 0$), X et Y sont donc dépendantes mais pas corrélées.

En résumé, l'indépendance implique l'absence de corrélation, mais la dépendance n'implique pas les corrélations (exemple ci-dessus). Et plus généralement, corrélation n'implique pas causalité...

Remarquons que pour un nombre de variables aléatoires supérieur à deux, les variables peuvent être indépendantes par paire, mais pas forcément indépendantes mutuellement.

[16] D'après l'équation 1.4, $Var(X^* \pm Y^*) = Var(X^*) + Var(Y^*) \pm 2cov(X^*, Y^*) = 2(1 \pm r(X,Y)) \geq 0$ donc $|r| \leq 1$. Si $r = 1$, alors $X^* - Y^* = cte$ donc $Y = aX + b$, où a et b sont des constantes.

Exemple : Considérons l'ensemble des 6 permutations des trois lettres $l = a$, $l = b$ et $l = c$ complété par les trois événements $\{aaa\}$, $\{bbb\}$ et $\{ccc\}$. Soit $X_k = l$ l'événement "la lettre l est à la k^{eme} place", où $k = 1, 2, 3$. Alors $Pr(X_1 = a) = Pr(X_2 = b) = Pr(X_3 = c) = 1/3$ et $Pr(X_1 = a \; et \; X_2 = b) = Pr(X_1 = a).Pr(X_2 = b) = 1/9$. Les variables aléatoires X_1 et X_2 sont donc bien indépendantes. Par contre X_1, X_2 et X_3 ne le sont pas, car $Pr(X_1 = a; X_2 = b; X_3 = c) = 1/9 \neq Pr(X_1 = a)Pr(X_2 = b)Pr(X_3 = c)$.

1.2.5 Fonction caractéristique

Nous verrons par la suite que la fonction caractéristique est un outil puissant pour étudier une distribution de probabilité. On appelle fonction caractéristique la transformée de Fourier $\phi(u) = \hat{P}(u) = \langle e^{iuX} \rangle$ de la densité de probabilité $P(x)$. Ainsi

$$\phi(u) = \langle e^{iuX} \rangle = \sum_j e^{iux_j} P(x_j) \quad \text{ou} \quad \phi(u) = \int e^{iux} P(x) dx.$$

Une distribution est donc entièrement déterminée par sa fonction caractéristique. Donnons quelques exemples d'utilisation de la fonction caractéristique.

- Calcul des moments de la distribution. En effet,

$$\langle X^n \rangle = \frac{1}{i^n} \frac{d^n \phi(u)}{du^n} \bigg|_{u=0}. \tag{1.5}$$

  Une autre méthode pour calculer les moments est de développer $\phi(u)$ en puissance de u.

- Le *cumulant* d'ordre n est lui défini par

$$c_n = (-i)^n \frac{d^n \ln \phi(u)}{du^n} \bigg|_{u=0}. \tag{1.6}$$

  On montre facilement que $c_1 = \langle X \rangle$ et $c_2 = Var(X)$. En général, le cumulant d'ordre n s'exprime en fonction des moments d'ordre $p \leq n$. On introduit en particulier la *kurtosis* (coefficient d'aplatissement) :

$$\kappa = \frac{c_4}{c_2} = \frac{\langle (X - \langle X \rangle)^4 \rangle}{\sigma^2} - 3.$$

  Cette quantité mesure l'aplatissement d'une distribution et la déviation par rapport à la distribution gaussienne. Ainsi, une valeur de κ élevée caractérise une distribution piquée autour de sa moyenne avec des queues épaisses (probabilité "élevée" d'événements extrêmes).[17]

[17] Nous verrons que pour la distribution gaussienne, $\kappa = 0$. Si $\kappa > 0$, la distribution est *leptokurtique*, si $\kappa < 0$, la distribution est *platikurtique*.

- Produit de convolution. Soient deux variables aléatoires X et Y, de densité de probabilité $P(x)$ et $Q(x)$. Exprimons la densité de probabilité $H(s)ds$ de la variable aléatoire $S = X + Y$.
  La transformée de Fourier de $H(s)$ est :

$$\hat{H}(u) = \langle e^{iu(X+Y)} \rangle = \langle e^{iuX}.e^{iuY} \rangle.$$

  Par ailleurs, si X et Y sont indépendantes

$$\langle e^{iuX}.e^{iuY} \rangle = \langle e^{iuX} \rangle \langle e^{iuY} \rangle = \hat{P}(u)\hat{Q}(u).$$

  En appliquant la transformée de Fourier inverse et le théorème sur la transformée de Fourier du produit de convolution,[18] on retrouve l'équation (1.3) :

$$H(s) = P * Q = \int P(s-t)Q(t)dt.$$

- Somme de variables aléatoires indépendantes. Soient N variables aléatoires X_j, où $j = 1, 2, ...N$, *indépendantes* de distribution $P_j(x)$ dont la fonction caractéristique est $\phi_j(u)$. En utilisant à nouveau le produit de convolution, on montre que la fonction caractéristique de la variable $S_N = \sum_{j=1}^{N} X_j$ est :

$$\phi_{S_N}(u) = \langle e^{iuS_N} \rangle = \prod_{j=1}^{N} \phi_j(u).$$

  Dans le cas particulier où les N variables ont la même distribution, on a donc

$$\phi_{S_N}(u) = \phi_j^N(u).$$

Cas de variables aléatoires discrètes : fonction génératrice

Soit X une variable aléatoire pouvant prendre les valeurs $k = 0, 1, 2, \cdots$ avec des probabilités $P(k)$. C'est par exemple le cas de la distribution binomiale ou de la loi de Poisson (voir section 1.3). Plutôt que d'utiliser la fonction caractéristique en introduisant la fonction de Dirac, on peut définir[19] une *fonction génératrice* $A(s)$:

$$A(s) = \langle s^X \rangle = \sum_k P(k)s^k.$$

On voit que $A(s)$ est la valeur moyenne d'une nouvelle variable aléatoire s^X. On a immédiatement $A(1) = 1$, $A'(1) = \langle X \rangle$, $A''(1) = \langle X(X-1) \rangle$, etc. On montre aussi facilement que si $A(s)$ et $B(s)$ sont fonctions génératrices des variables aléatoires *indépendantes* X et Y, alors $C(s) = A(s)B(s)$ est fonction génératrice de la variable aléatoire $S = X + Y$.

[18] La transformée de Fourier d'un produit de convolution est le produit des transformées de Fourier :

$$\hat{H}(u) = \int e^{ius}H(s)ds = \int e^{ius}ds \int P(s-t)Q(t)dt = \int dte^{iut}Q(t) \int dse^{iu(s-t)}P(s-t) = \hat{P}(u)\hat{Q}(u).$$

[19] En mécanique statistique, c'est la grande fonction de partition.

1.3 Quelques distributions de probabilité

Seulement quelques lois de probabilités interviennent fréquemment dans les phénomènes naturels et les sciences sociales. Il s'agit en particulier des lois binomiale, de Poisson et gaussienne dont les relations asymptotiques sont résumées sur la figure 1.1.

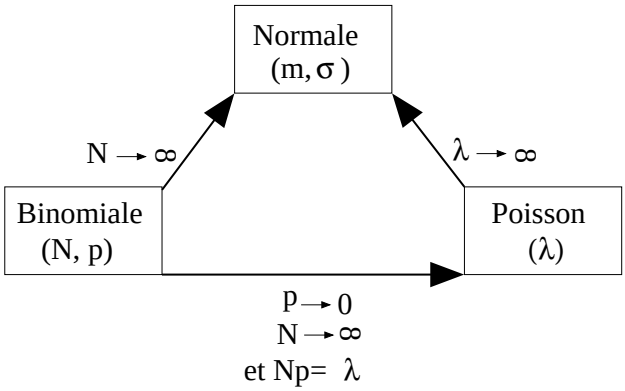

FIG. 1.1 – *Relations asymptotiques entre les distributions gaussienne, binomiale et de Poisson.*

distribution	variable	paramètres	fonction	valeur moyenne	variance
Uniforme	$a \leq x \leq b$	a,b réels	$P(x) = \frac{1}{b-a}$	$m = (b+a)/2$	$(b-a)^2/12$
Gaussienne	x réel	m réel, σ réel positif	$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$	m	σ^2
Binomiale	$k = 0, 1, ..., N$ entier	N entier, $0 \leq p \leq 1$	$P(k) = C_N^k p^k (1-p)^{N-k}$	Np	$Np(1-p)$
Poisson	$k = 0, 1, ...$ entier	λ réel positif	$P(k) = e^{-\lambda} \frac{\lambda^k}{k!}$	λ	λ
χ^2	x réel positif	N entier positif	$P(x) = \frac{1}{2^{\frac{N}{2}}} \frac{x^{\frac{N}{2}-1}}{\Gamma(\frac{N}{2})} e^{-\frac{x}{2}}$	N	$2N$

1.3.1 Distribution uniforme

C'est la distribution qui résulte d'une expérience dans laquelle chaque évé-nement a la même probabilité. On dit que les événements sont *équiprobables.*[20] Par exemple, lorsqu'on lance un dé parfait, chaque face a une probabilité 1/6 d'apparaître.

Soit une variable aléatoire réelle[21] X dont les valeurs sont comprises entre a et b. La distribution de probabilité uniforme est donnée par

$$P(x) = \frac{1}{b-a} \quad si \quad a \leq x \leq b, \quad 0 \quad sinon. \tag{1.7}$$

Sa valeur moyenne est $\langle X \rangle = (b+a)/2$ et sa variance $Var(X) = (b-a)^2/12$. Dans le cas d'une variable entière, ces formules sont fausses en général (voir l'exercice 4).

Exercice 4 *: Lancé de dé* †

Le chiffre N_1 obtenu par un lancé de dé est distribué uniformément avec une probabilité 1/6*. Calculer la valeur moyenne et la variance de N_1.*

Exercice 5 *: Somme de deux variables aléatoires uniformes* ‡

Le lancé de deux dés fournit deux nombres aléatoires N_1 et N_2. Quelle est la distribution de la somme $N = N_1 + N_2$? Calculer la valeur moyenne et la variance de N.

1.3.2 Loi gaussienne ou normale

Nous verrons au chapitre suivant pourquoi cette loi de probabilité est si fréquente. Soit une variable aléatoire X continue qui varie sur $\mathcal{R}$ tout entier. La distribution gaussienne[22] (ou normale) dépend de deux paramètres réels m et $\sigma > 0$:

$$\boxed{P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}} \tag{1.8}$$

[20] En physique statistique, c'est la distribution de probabilité des états microscopiques d'un système isolé (ensemble *microcanonique*).

[21] La variable peut également être discrète comme dans le cas du lancé de dé.

† **Réponse** : On trouve facilement en examinant les 6 événements possibles : $\langle N_1 \rangle = 3.5$ et $Var(N_1) = 35/12$. Attention, la formule de la variance d'une variable réelle aurait donné un résultat faux (25/12).

‡ **Réponse** : La probabilité $p(n)$ d'obtenir une somme égale à $n = 2, ...12$ est

$$p(n) = (1/6)^2 \sum_{n_1=1}^{6} \sum_{n_2=1}^{6} \delta(n - n_1 - n_2).$$

En énumérant tous les cas possibles, on trouve : $p(n) = (n-1)/36$ si $2 \leq n \leq 7$ et $p(n) = (13-n)/36$ si $8 \leq n \leq 12$. La somme de deux variables distribuées uniformément n'est donc pas distribuée uniformément : on dit que cette distribution n'est pas *stable* (voir chapitre 1.4). Puisque N_1 et N_2 sont indépendantes, on retrouve facilement $\langle N \rangle = 7$ et $Var(N) = 35/6$.

[22] Carl Friedrich Gauss (1777-1855) mathématicien et physicien allemand.

Cette loi est bien normalisée ($\int_{-\infty}^{+\infty} P(x)\,dx = 1$) et sa valeur moyenne et sa variance sont données[23] par $\langle X\rangle = m$ et $Var(X) = \sigma^2$. La valeur moyenne est donc la valeur la plus probable. La fonction caractéristique de la loi normale est

$$\phi(u) = e^{ium - u^2\sigma^2/2}.$$

Ce qui permet, par exemple, de calculer les moments de la distribution en utilisant la relation (1.5). En utilisant la définition (1.6), on remarque une spécificité de la gaussienne : les cumulants c_n sont nuls pour $n > 2$. En particulier sa kurtosis est nulle.

Ainsi, on sait depuis le XIX^{eme} siècle que la taille des individus est distribuée sur une loi gaussienne décrite uniquement par sa moyenne m et sa variance σ^2. Par exemple, pour les japonais, $m = 165.5$ cm et $\sigma = 5.8$ cm, et pour les américains, $m = 175.5$ cm et $\sigma = 7.1$ cm. Il est intéressant de garder en tête les valeurs des probabilités d'écart à la moyenne :

$$Pr(m - \sigma < X < m + \sigma) \simeq 68\%$$

et

$$Pr(m - 2\sigma < X < m + 2\sigma) \simeq 96\%.$$

Ces probabilités se "lisent" clairement sur la figure 1.3 comme l'aire sous la courbe avec une base de largeur 2σ et 4σ respectivement. On voit que contrairement à une loi large (voie section 1.3.6) une variable gaussienne s'écarte peu de sa moyenne.

Quand une variable aléatoire a une densité de probabilité gaussienne, on dit qu'elle est *normale*, si elle a une moyenne nulle elle est *centrée* et si sa variance vaut 1, elle est *réduite*.[24]

[23] En effet, Il faut calculer

$$\langle X\rangle = \int xP(x)\,dx = \frac{1}{\sqrt{\pi}}[\int_{-\infty}^{\infty} \sigma\sqrt{2}te^{-t^2}\,dt + m\int_{-\infty}^{\infty} e^{-t^2}\,dt],$$

où l'on a effectuer le changement de variable $t = \frac{x-m}{\sigma\sqrt{2}}$. La première intégrale est nulle par parité et la seconde vaut $\sqrt{\pi}$ (voir l'annexe A.2). Calculons maintenant

$$\langle (X - \langle X\rangle)^2\rangle = \int (x-m)^2 P(x)\,dx.$$

Posons

$$J(\sigma) = \int e^{-\frac{(x-m)^2}{2\sigma^2}}\,dx = \sqrt{2\pi}\sigma$$

et dérivons par rapport à σ

$$\frac{dJ}{d\sigma} = \frac{1}{\sigma^3}\int (x-m)^2 e^{-\frac{(x-m)^2}{2\sigma^2}}\,dx,$$

d'où on tire immédiatement

$$\langle (X-m)^2\rangle = \sigma^2.$$

[24] Voir la définition d'une variable réduite section 1.2.2.

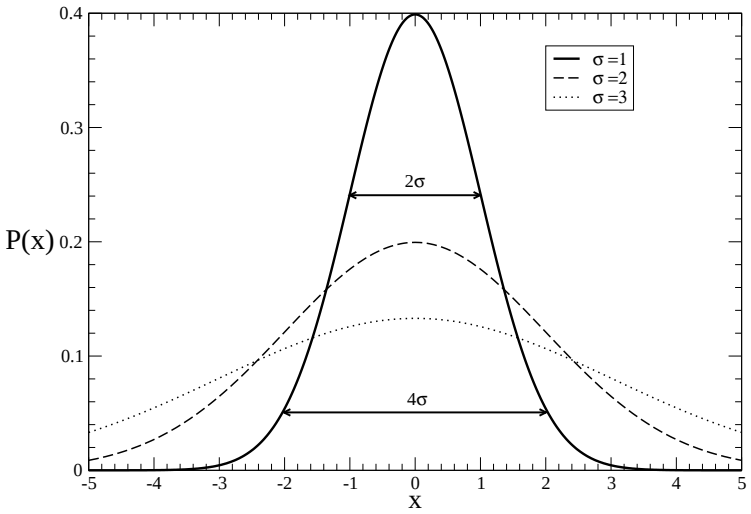

FIG. 1.2 – *Distributions normales centrées pour* $\sigma = 1, 2$ *et 3.*

Loi log-normale

Si une variable aléatoire X suit une loi gaussienne de moyenne m et de variance σ^2, la variable $Y = e^X$ est distribuée selon une loi dite *log-normale* :

$$Q(y) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{y} e^{-\frac{(\ln y - m)^2}{2\sigma^2}} \quad \text{pour} \quad y > 0.$$

La fonction caractéristique n'est pas définie.

Exercice 6 *: Loi log-normale*[†]
Démontrer ce résultat et montrer que $\langle Y \rangle = e^{m+\sigma^2/2}$ *et* $Var(Y) = e^{2m+\sigma^2}(e^{\sigma^2} - 1)$. *Montrer que la valeur la plus probable est* $Y_m = e^{m-\sigma^2} < \langle Y \rangle$.

Nous verrons à la section 1.4 que la distribution du produit d'un grand nombre de variables aléatoires indépendantes tend vers une loi log-normale.

Cette distribution joue un rôle fondamental en finance, puisque le modèle de Black-Scholes suppose que la variation relative des valeurs boursières $\delta \ln S$ est distribuée selon une loi normale.

[†]**Réponse** : Il s'agit d'une composition de variables aléatoires avec $Y = F(X) = e^X$ (voir section 1.2.2). Les calculs de $\langle Y \rangle$ et $Var(Y)$ ne présentent pas de difficulté. Si la variable X est certaine ($\sigma = 0$), on trouve bien $Var(Y) = 0$.

.

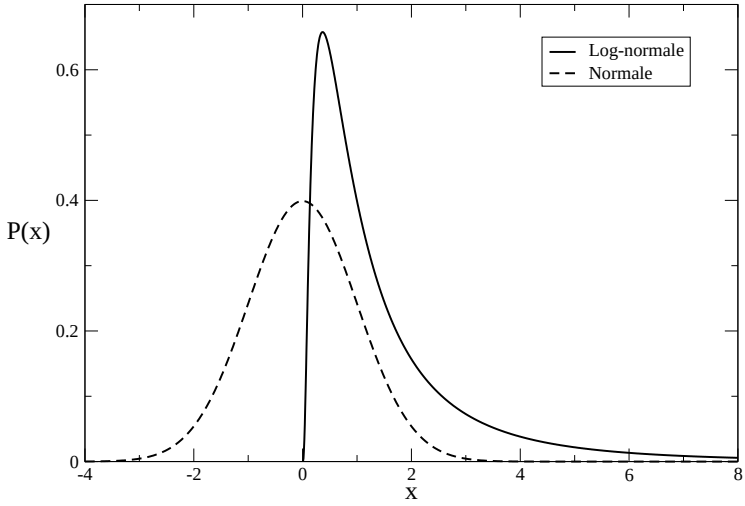

FIG. 1.3 – *Distributions normale et log-normale pour $m = 0$ et $\sigma = 1$.*

1.3.3 Loi binomiale

Cette loi intervient lorsque l'on répète *indépendamment* un événement qui n'a que deux issues possibles de probabilité p et $q = 1 - p$ respectivement ("Bernouilli trials" en anglais). Par exemple, le jeu de "pile ou face" avec une pièce parfaite implique $p = q = 1/2$. La loi binomiale est la probabilité de réaliser k fois exactement un événement de probabilité p en effectuant N essais *indépendants* (la variable aléatoire entière K varie donc de 0 à N) :

$$\boxed{P(k) = C_N^k p^k (1-p)^{N-k}} \tag{1.9}$$

où C_N^k est le nombre de façons de choisir k objets parmi N, donné par l'équation (1.1). On montre facilement[†] que $\sum_{k=0}^{N} P(k) = 1$ (normalisation), $\langle K \rangle = Np$ et $Var(K) = Np(1-p)$. La fonction caractéristique de la loi binomiale est $\phi(u) = (pe^{iu} + 1 - p)^N$.

Exercice 7 *: Valeur la plus probable* [*]

Montrer que la valeur moyenne est (presque) la valeur la plus probable (maximum de la fonction).

[†] A l'aide de la formule du binôme de Newton que l'on peut dériver par rapport à p : $(p+q)^N = \sum_{k=0}^{N} C_N^k p^k q^{N-k}$.

[*] **Réponse** : Pour montrer que le maximum est en Np, on peut calculer le rapport

$$\frac{P(k)}{P(k-1)} = 1 + \frac{(N+1)p - k}{k(1-p)}$$

qui est supérieur à 1 pour $k\langle(N+1)p$ et inférieur pour $k > (N+1)p$.

Exercice 8 *: Lapins albinos* †

Un lapin sur 100 est albinos. On prend aléatoirement 100 lapins dans un clapier, quelle est la probabilité qu'il y en ait un (exactement) albinos ?

Quand $N \to \infty$, la distribution binomiale (1.9) tend vers la loi gaussienne (pour p pas trop proche de 0 ou 1)(1.8)§, où $k \in [0, N]$ devient une variable continue $x \in [0, +\infty]$, avec $\langle X \rangle = pN$ et $\sigma^2 = Np(1-p)$. Cette propriété de la loi binomiale est illustrée par la figure 1.4.

Loi multinomiale

C'est une généralisation de la loi binomiale. Une variable aléatoire K peut prendre r valeurs $(k_1, k_2, \cdots k_r)$ avec les probabilités respectives $p_1, p_2, \cdots p_r$ telle que $\sum_{i=1}^{r} p_i = 1$. La probabilité d'obtenir k_i tirages (indépendants) de $K = k_i$ en N essais ($\sum_{i=1}^{r} k_i = N$) est donnée par

$$P(k_1, k_2 \cdots k_r) = \frac{N!}{k_1! k_2! \cdots k_r!} p_1^{k_1} p_2^{k_2} \cdots p_r^{k_r}.$$

La démonstration est évidente : $p_1^{k_1} p_2^{k_2} \cdots p_r^{k_r}$ est la probabilité d'obtenir le résultat dans un ordre donné et la dégénérescence vaut

$$\frac{N!}{k_1! k_2! \cdots k_r!} = C_N^{k_1} C_{N-k_1}^{k_2} \cdots C_{N-k_1-k_2\cdots-k_{r-1}}^{k_r}.$$

C'est la distribution de probabilité obtenue lors de tirages indépendants de boules de r couleurs différentes contenues dans une urne. Les valeurs moyennes, les variances et les covariances sont données respectivement par $\langle K_i \rangle = np_i$, $Var(K_i) = Np_i(1-p_i)$ et $cov(K_i, K_j) = -Np_i p_j$.

---

†**Réponse** : La probabilité pour que le *premier* lapin choisi soit albinos et qu'aucun des 99 suivants ne le soit est 1/100 multiplié par $(99/100)^{99}$ (les probabilités indépendantes se multiplient). Maintenant on ne demande pas qu'un lapin déterminé (le premier) soit albinos. Ce peut être le second, le troisième ou ...le dernier. Il y a 100 telles possibilités. Comme elles s'excluent mutuellement, la probabilité qu'un lapin *quelconque* soit albinos est donc 100 fois la probabilité précédente, soit $(99/100)^{99} \simeq 1/e \simeq 0.37$.

On aurait pu aussi partir directement de la loi binomiale (1.9) et écrire que $P(1) = C_{100}^1 \frac{1}{100} \left(1 - \frac{1}{100}\right)^{99}$. "Un lapin sur 100 est albinos" ne signifie donc pas que sur 100 lapins tirés au hasard il y en a nécessairement un qui le soit. Par contre, si on effectue une série de 100 prélèvements dans le clapier, on aura bien *en moyenne* $Np = 100/100 = 1$ lapin sur 100 albinos.

§Pour le montrer, on peut utiliser la fonction caractéristique de la distribution binomiale $\phi_b(u)$ et montrer qu'elle tend vers la fonction caractéristique de la loi normale $\phi_n(u)$. En effet,

$$\begin{aligned}
\ln \phi_b(u) &= N \ln(1 + p(e^{iu} - 1)) = N[p(e^{iu} - 1) - \frac{p^2}{2}(e^{iu} - 1)^2 + ...] \\
&= N[p(iu - \frac{u^2}{2} + ...) - \frac{p^2}{2}(iu - \frac{u^2}{2} + ...)^2 + ...] \\
&\simeq iuNp - \frac{u^2}{2} Np(1-p) = \ln \phi_n(u).
\end{aligned}$$

.

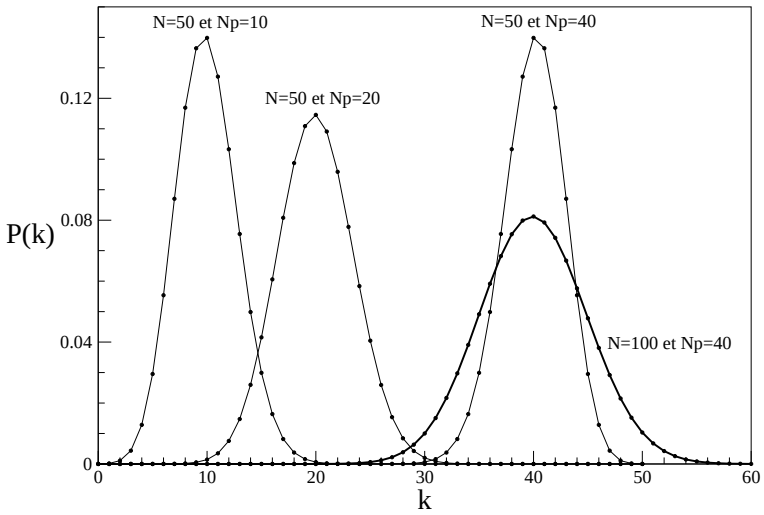

FIG. 1.4 – *Distributions binomiales pour différentes valeurs de N et de Np. Contrairement aux trois courbes calculées pour $N = 50$, la distribution obtenue pour $N = 100$ et $Np = 40$ est indiscernable d'une loi gaussienne (avec $m = 40$ et $\sigma^2 = 24$).*

1.3.4 Loi de Poisson

Lorsque des événements *indépendants* se produisent aléatoirement avec un taux constant (nombre d'événements par unité de temps ou d'espace), la loi de Poisson[25] est la probabilité d'en trouver $k = 0, 1, 2, ...$ alors que l'on en attend λ en moyenne :

$$\boxed{P(k) = e^{-\lambda}\lambda^k/k!} \tag{1.10}$$

On montre facilement que la loi de Poisson est normalisée et que la valeur moyenne de K et sa variance sont égales : $\langle K\rangle = Var(K) = \lambda$. La fonction caractéristique de la loi de Poisson est $\phi(u) = e^{\lambda(e^{iu}-1)}$.

La loi de Poisson est la limite de la loi binomiale (1.9) quand $N \to \infty$ (nombreux essais) et $p \to 0$ (événement rare) de telle sorte que $Np = \lambda$ soit fini.¶ Par ailleurs, la loi de Poisson (1.10) tend, lorsque $\lambda \to \infty$, vers la loi

[25] Siméon Denis Poisson (1781-1840) mathématicien français.

¶ A l'aide des fonctions caractéristiques $\phi_b(u)$ et $\phi_p(u)$ des distributions binomiale et de Poisson quand $N \to \infty$:

$$\begin{aligned}\ln\phi_b(u) &= N\ln(1+p(e^{iu}-1)) = (Np)(e^{iu}-1) - \frac{(Np)^2}{2N}(e^{iu}-1)^2 + ...] \\ &\simeq Np(e^{iu}-1) = \ln\phi_p(u).\end{aligned}$$

gaussi

.

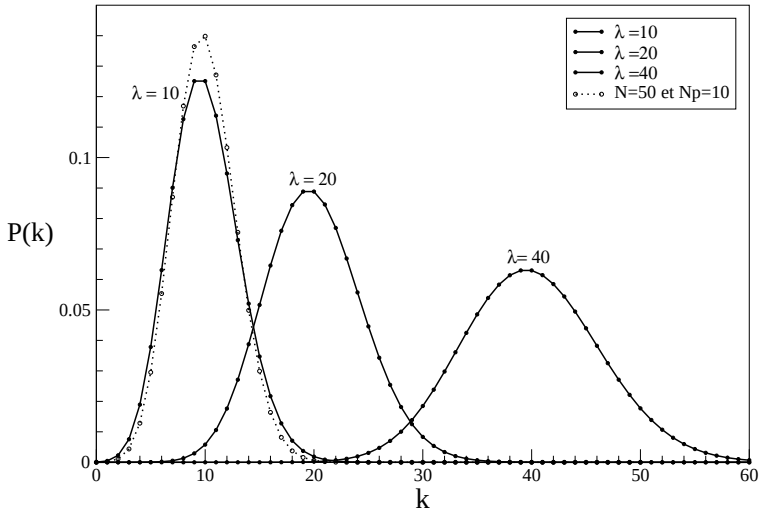

FIG. 1.5 – *Distributions de Poisson pour différentes valeurs de λ. La distribution binomiale ($N = 50$ et $Np = 10$) est représentée en pointillés à titre de comparaison. La loi de Poisson ($\lambda = 10$) est indiscernable de la loi binomiale ($N = 1000$ et $Np = 10$). De même la loi de Poisson ($\lambda = 40$) est indiscernable d'une distribution gaussienne ($m = \sigma^2 = 40$).*

Exercice 9 *: Anniversaire* [§]

Quelle est la probabilité pour que dans un amphi de $N = 500$ étudiants, exactement k personnes aient leur anniversaire le 16 janvier ?

Exercice 10 *: Comparaison entre lois* [¶]

Quelle est la probabilité d'obtenir k fois un 6 en N lancés de dés ?

[‡]Une fois de plus :

$$\ln \phi_p(u) = \lambda(e^{iu} - 1) = \lambda[iu - \frac{u^2}{2} + ...] \simeq \ln \phi_p(u).$$

[§]**Réponse** : La probabilité qu'une personne ait son anniversaire à une date donnée est $p = 1/365$. La probabilité recherchée est la distribution de Poisson en prenant $\lambda = 500/365$. Comparer ce résultat pour $k = 0, 1, 2, 3$ avec la probabilité obtenue avec la distribution binomiale.

[¶]**Réponse** : Cette probabilité est donnée par la loi binomiale

$$P(k) = C_N^k \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{N-k}.$$

Sur la figure 1.6 on a tracé l'histogramme pour $N = 10$ et comparé aux approximations gaussienne et de Poisson. Aucune n'est très bonne (ni très mauvaise !).

.

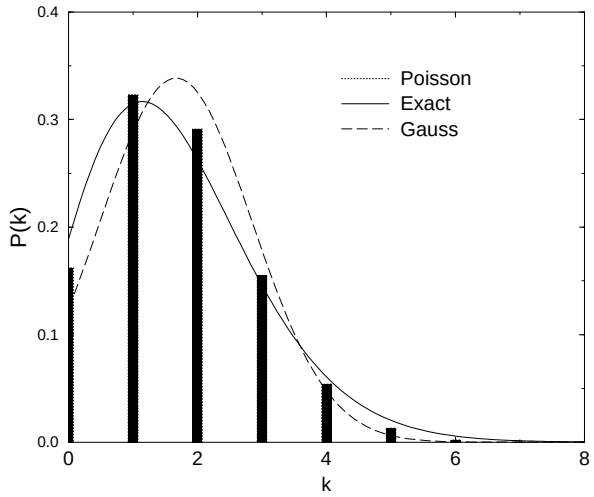

FIG. 1.6 – *Comparaison des lois de probabilité pour* $p = 1/6$ *et* $N = 10$. *Le calcul exact est donné par la loi binomiale.*

Processus de Poisson

Comme nous le verrons par la suite, le processus de Poisson est une bonne illustration d'un processus stochastique.

Considérons par exemple N appels téléphoniques[26] passés au hasard et indépendamment pendant un temps T. On s'attend donc à recevoir un appel toutes les T/N *en moyenne*. Comme on le voit sur la figure 1.7, la probabilité qu'un appel soit passé pendant une durée $t \leq T$ est $p = t/T$, quelle que soit l'origine de t (autrement dit, la probabilité qu'un appel ne soit pas donné pendant la durée t est $1 - t/T$).

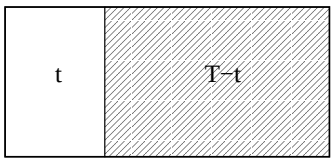

FIG. 1.7 – *Le rectangle de surface* T *est divisé en deux régions de surface* t *et* $T - t$. *Sur les* N *appels qui vont "tomber" dans le rectangle, une fraction* t/T *touchera la région de surface* t.

[26] Mais nous aurions pu étudier le nombre de particules α émises par un matériau radioactif pendant un intervalle de temps donné ou la distribution spatiale des impacts de bombes tombées sur Londres pendant la seconde guerre mondiale.

Soient $t_1, t_2, \cdots, t_N$, les dates des premier, second ... et dernier appels. Ce sont des variables aléatoires.

• Pour commencer, considérons une autre variable aléatoire $\Delta t_i = t_{i+1} - t_i$, l'intervalle de temps entre le i^{eme} appel et le suivant. Puisque les appels sont indépendants, la distribution $P(\tau)$ du temps d'attente entre deux appels ne dépend pas de i. En raisonnant sur la répartition temporelle des N appels, l'événement $\Delta t_i \in [\tau, \tau + d\tau]$ est la réunion de deux événements indépendants : sur N appels, aucun n'arrive pendant le temps τ (probabilité que $N-1$ appels n'aient pas lieu pendant le temps τ) et un appel arrive pendant le temps $d\tau$ (il ne peut pas en arriver plusieurs si $d\tau \ll T/N$). On a donc

$$Pr\left(\Delta t_i \in [\tau, \tau + d\tau]\right) = P(\tau)d\tau = C_N^1 \left(1 - \frac{\tau}{T}\right)^{N-1} \frac{d\tau}{T}.$$

On suppose maintenant que $T \to \infty$ et $N \to \infty$ de telle sorte que $N/T = \lambda$ reste constant (taux moyen d'appels ou nombre d'appels par unité de temps). On en tire par passage à la limite que

$$P(\tau) = \lambda e^{-\lambda\tau}.$$

C'est une *distribution exponentielle* de valeur moyenne $1/\lambda = T/N$ et de variance $1/\lambda^2$ (il est étonnant que le maximum de P soit obtenu pour $\tau = 0$). Cette densité de probabilité est fréquemment utilisée pour décrire des durées de vie, ou les intervalles de temps entre les arrivées successives de clients à une file d'attente.[27]

• Introduisons une nouvelle variable aléatoire K_t, le nombre d'appels passés pendant la durée t, qui va nous permettre de calculer la loi (discrète) de distribution $P_t(k)$ du nombre d'appels pendant une durée t. Comme précédemment on peut écrire,

$$Pr(K_t = k) = P_t(k) = C_N^k \left(1 - \frac{t}{T}\right)^{N-k} \left(\frac{t}{T}\right)^k.$$

À la limite $N \to \infty$, $p = t/T = t\lambda/N \to 0$, on est bien dans les conditions de l'approximation de Poisson et on retrouve[28] la distribution du même nom :

[27] Cette distribution a une propriété évidente de "perte de mémoire" :

$$Pr(X > \tau_1 + \tau_2) = \int_{\tau_1+\tau_2}^{\infty} P(\tau)d\tau = Pr(X > \tau_1)Pr(X > \tau_2)$$

.

[28] On le voit également directement :

$$\begin{aligned} P_t(k) &= \frac{N!}{k!(N-k)!}\left(1 - \frac{t\lambda}{N}\right)^{N-k}\left(\frac{t\lambda}{N}\right)^k \simeq \frac{N^k}{k!}\left(1 - \frac{t\lambda}{N}\right)^{N-k}\left(\frac{t\lambda}{N}\right)^k \\ &\simeq \frac{(\lambda t)^k}{k!}\left(1 - \frac{t\lambda}{N}\right)^N \simeq \frac{(\lambda t)^k}{k!}e^{-\lambda t}. \end{aligned}$$

$$P_t(k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}. \tag{1.11}$$

Le nombre moyen d'appels pendant une durée t est donc $\langle K_t \rangle = \lambda t = Nt/T$.

• On peut enfin chercher les lois (continues) de distribution $f_i(t_i)$ des t_i (date du i^{eme} appel). Le plus simple est d'écrire que

$$t_i = t_{i-1} + \tau,$$

où t_{i-1} et τ sont indépendants. Donc

$$f_i(t_i) = f_{i-1}(t_{i-1}) * P(\tau).$$

Et comme, manifestement, $f_1(t_1) = P(t_1) = \lambda e^{-\lambda t_1}$, on calcule (sachant que $t_1 \leq t_2$)

$$f_2(t_2) = P(t_1) * P(\tau) = \int_0^{t_2} P(t_1) P(t_2 - t_1) dt_1 = \lambda^2 t_2 e^{-\lambda t_2}.$$

Par récurrence,[29] on en déduit que

$$f_i(t) = \frac{\lambda^i t^{i-1}}{(i-1)!} e^{-\lambda t}. \tag{1.12}$$

On a bien $\int_0^\infty f_i(t)dt = 1$ et en moyenne le i^{eme} appel a lieu à l'instant $i/\lambda = iT/N$ (voir la définition des fonctions Gamma dans l'annexe A.2). La distribution $f_i(t)$ est appelée *loi Gamma* (voir section 1.3.5)

La similitude entre les équations (1.11) et (1.12) n'est pas surprenante : L'événement "le nombre d'appels K_t pendant le temps t est supérieur à n" est le même que "t_n est inférieur à t". Comme l'événement $K_t \geq n$ est la réunion des deux événements disjoints $K_t \geq n+1$ et $K_t = n$, on a l'égalité

$$Pr(K_t = n) = Pr(K_t \geq n) - Pr(K_t \geq n+1).$$

Elle se traduit par

$$Pr(K_t = n) = Pr(t_n \leq t) - Pr(t_{n+1} \leq t).$$

Il reste alors à calculer le deuxième membre. Il vaut

$$\int_0^{\lambda t} \left[\frac{e^{-u} u^{n-1}}{(n-1)!} - \frac{e^{-u} u^n}{n!} \right] du = \int_0^{\lambda t} \frac{d}{du} \frac{e^{-u} u^n}{n!} du = e^{-\lambda t} \frac{(\lambda t)^n}{n!}.$$

[29] En effet, supposons que nous connaissons l'expression de $f_i(t_i)$ donnée par l'équation (1.12) :

$$f_{i+1}(t) = \int_0^t f_i(t_i) f_1(t - t_i) dt_i = \int_0^t \frac{\lambda^i t_i^{i-1}}{(i-1)!} e^{-\lambda t_i} . \lambda e^{-\lambda(t-t_i)} dt_i = \frac{\lambda^{i+1} t^i}{i!} e^{-\lambda t}.$$

1.3.5 Loi Gamma et loi du χ^2

Loi Gamma

La loi Gamma, que nous avons rencontrée en étudiant le processus de Poisson, est typique des processus de file d'attente :

$$P(x) = \frac{\lambda}{\Gamma(\nu)}(\lambda x)^{\nu-1}e^{-\lambda x}.$$

où λ et ν sont deux réels positifs et $\Gamma(\nu) = (\nu-1)!$ est la fonction Gamma (voir annexe A.2). On peut montrer que sa moyenne, sa variance et sa fonction caractéristique sont données par $\langle X\rangle = \nu/\lambda$, $Var(X) = \nu/\lambda^2$ et $\phi(u) = (1 - i\frac{u}{\lambda})^{-\nu}$.

Un exemple d'utilisation de la loi Gamma est donné par la figure 1.8.

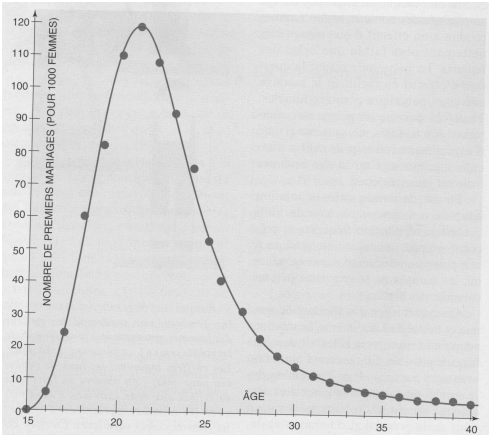

FIG. 1.8 – *Distribution de l'âge auquel les Françaises nées dans les années 1930 se sont mariées pour la première fois (Dossier Pour la Science "Le hasard", Avril 1996, p46). Cette distribution est ajustée par une loi Gamma de paramètres* $\lambda \simeq 0.57$ *ans*$^{-1}$ *et* $\nu \simeq 4.42$ *et translatée de 15 ans.*

La *loi exponentielle*, $P(x) = \lambda e^{-\lambda x}$, est un cas particulier de la loi Gamma (pour $\nu = 1$). Sa fonction caractéristique est donc $\phi(u) = (1 - i\frac{u}{\lambda})^{-1}$. Ainsi, lorsque la densité de probabilité des intervalles de temps entre deux événements est une loi exponentielle de paramètre λ, la loi gamma (λ, ν) est la distribution des intervalles de temps entre le i^{eme} et le $(i+\nu)^{eme}$ événement.

Plus généralement, comme on le voit à l'aide des fonctions caractéristiques, pour $\nu = N$, la loi Gamma est la distribution de la somme de N variables aléatoires *indépendantes* distribuées sur une loi exponentielle de même paramètre λ. Et une fois de plus, lorsque $N \to \infty$, la distribution Gamma (λ, N) tend vers une loi normale de moyenne $m = N/\lambda$ et de variance $\sigma^2 = N/\lambda^2$.

Loi du χ^2

La loi du χ^2 joue un rôle important en statistique, en particulier pour comparer des données expérimentales avec une loi théorique. Soit $X = \chi^2$ une variable aléatoire réelle et positive. La distribution du χ^2 dépend d'un paramètre entier positif, N, appelé le nombre de degrés de liberté ; elle est donnée par :

$$P(x) = \frac{1}{2^{\frac{N}{2}}} \frac{x^{\frac{N}{2}-1}}{\Gamma(\frac{N}{2})} e^{-\frac{x}{2}},$$

où Γ est une fonction Gamma (voir l'annexe A.2). Clairement, la loi du χ^2 est un cas particulier de la loi Gamma (pour $\lambda = 1/2$ et $\nu = N/2$). On calcule facilement $\langle X \rangle = N$, $\langle X^2 \rangle = N(N+2)$ et $Var(X) = 2N$. Sa fonction caractéristique est $\phi(u) = (1-2iu)^{-N/2}$.

Illustrons son intérêt par la propriété suivante : soient $X_1, ..., X_N$, N variables aléatoires *indépendantes* distribuées selon une loi normale, réduite et centrée. Posons

$$\chi^2(N) = \sum_{i=1}^{N} X_i^2.$$

Montrons que cette nouvelle variable aléatoire est distribuée selon la loi du χ^2 à N degrés de liberté. Puisqu'il s'agit de variables indépendantes, la densité de probabilité des X_i est

$$Pr(X_1 = x_1, X_2 = x_2, ...X_N = x_N) = \frac{1}{\left(\sqrt{2\pi}\right)^N} e^{-\frac{x_1^2+x_2^2+\cdots+x_N^2}{2}}.$$

La probabilité pour que $\chi^2 \in [y, y+dy]$ s'écrit

$$Pr\left(\chi^2 \in [y, y+dy]\right) = P(y)dy = \frac{1}{\left(\sqrt{2\pi}\right)^N} \int_{\sum_{i=1}^N x_i^2 \in [y,y+dy]} e^{-\frac{x_1^2+x_2^2+\cdots+x_N^2}{2}} dx_1 dx_2 \cdots dx_n,$$

donc

$$P(y)dy = \frac{e^{-\frac{y}{2}}}{\left(\sqrt{2\pi}\right)^N} \int_{\sum_{i=1}^N x_i^2 \in [y,y+dy]} d\,x_1 d\,x_2 \cdots d\,x_n.$$

L'intégrale dans le second membre de cette équation est le volume compris entre les sphères de rayons $\sqrt{y}$ et $\sqrt{y+dy}$ dans un espace à N dimensions. Ce volume est donc égal à l'hypersurface $S_N(\sqrt{y})$ de la sphère de rayon $\sqrt{y}$ multipliée par l'épaisseur $\sqrt{y+dy} - \sqrt{y} \simeq dy/(2\sqrt{y})$. On a donc d'après l'expression de S_N calculée dans l'annexe A.4 :

$$P(y)dy = \frac{e^{-\frac{y}{2}}}{\left(\sqrt{2\pi}\right)^N} . S_N(\sqrt{y}) . \frac{1}{2\sqrt{y}} dy$$

où

$$S_N(\sqrt{y}) = N \frac{\pi^{N/2}}{\Gamma(N/2+1)} y^{\frac{N-1}{2}}.$$

Puisque $\Gamma(N/2+1) = (N/2)\Gamma(N/2)$, on en déduit que $P(y)$ est la distribution du χ^2. Dans le cas $N = 1$, on voit que le carré d'une variable aléatoire gaussienne est distribué selon la loi du χ^2 à 1 degré de liberté.

1.3.6 Lois larges et lois de puissance

Une distribution est dite *large* lorsqu'elle n'a pas de second moment fini (variance non définie). Les lois larges sont très fréquentes et en particulier en finance, mais comme nous le verrons, elles ne vérifient pas les hypothèses du théorème de la limite centrale (voir section 1.4).

Loi de Cauchy

La loi de Cauchy[30] (également appelée loi de Lorentz) est représentée sur la figure 1.9 :

$$P(x) = \frac{a}{\pi}\frac{1}{a^2 + (x - x_0)^2}. \tag{1.13}$$

Plus le paramètre $a > 0$ est grand, plus cette distribution, centrée en x_0, est étalée. La loi de Cauchy est bien normalisée, n'admet ni de valeur moyenne, ni de moment d'ordre plus élevé (loi large) et sa fonction caractéristique est[31] $\phi(u) = e^{ix_0 u - a|u|}$.

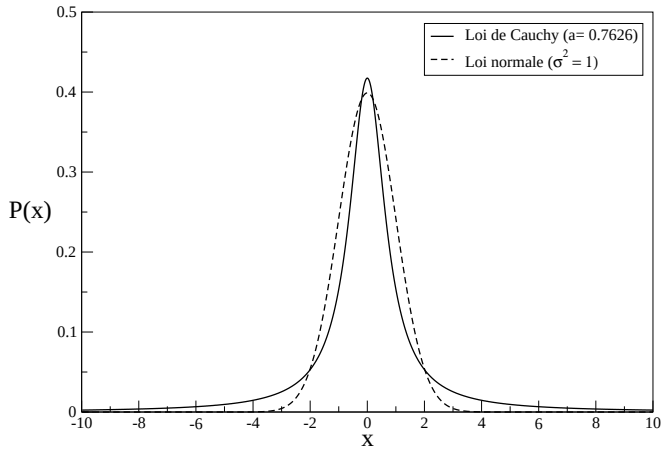

FIG. 1.9 – *Loi de Cauchy pour* $a = 0.7626$ *et loi gaussienne centrée et réduite.*

[30] Augustin Louis Cauchy (1789-1857) mathématicien français.

[31] On a bien

$$\frac{a}{\pi}\int_{-\infty}^{+\infty}\frac{1}{a^2 + (x - x_0)^2}dx = \frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{1}{1 + x^2}dx = \frac{1}{\pi}[\arctan(\infty) - \arctan(-\infty)] = 1.$$

La moyenne, $\langle X\rangle = \frac{a}{\pi}\int \frac{x}{a^2+(x-x_0)^2}dx$, n'existe pas puisque l'intégrand se comporte en $1/x$ pour $x \to \infty$. De même pour $\langle X^2\rangle$. La fonction caractéristique est :

$$\phi(u) = \frac{a}{\pi}\int_{-\infty}^{\infty} e^{iux}\frac{1}{a^2 + (x - x_0)^2}dx = e^{iux_0}e^{-a|u|}.$$

Les deux distributions représentées sur la figure 1.9 semblent très proches, mais le comportement asymptotique en loi de puissance ($\sim 1/x^2$ pour $x \to \infty$) de la loi de Cauchy permet des valeurs élevées de la variable aléatoire qui conduisent à des comportement très différents d'une loi gaussienne, comme nous le verrons au chapitre 2. Typiquement, le cours des actifs financiers est entre la loi de Cauchy (un grand écart à la moyenne correspondant à un Krach boursier) et la loi normale.

Lois de puissance

Les distributions en loi de puissance sont de la forme suivante :

$$P(x) = cx^{-\tau} \quad \text{pour} \quad x > x_0 > 0,$$

où c est une constante. Si l'exposant $\tau \leq 3$, ces lois sont larges (et si $\tau \leq 2$, non seulement la variance mais également la valeur moyenne divergent). Ce sont les seules lois qui n'ont pas de grandeur caractéristique, c'est-à-dire que leur allure ne dépend pas de *l'échelle* à laquelle on les étudie. Comme nous le verrons, les lois de puissances interviennent naturellement dans les phénomènes physiques où le système présente le même aspect quelle que soit l'échelle d'observation. C'est le cas en particulier des transitions de phase continues (point critique thermodynamique, seuil de percolation...).

Exercice 11 *: Loi de puissance* †
Montrer que si

$$P(bx) = g(b)P(x),$$

où b est un réel, alors $P(x)$ est une loi de puissance (la réciproque est évidente).

---

†**Réponse** :
On a

$$P(bx) = \frac{P(b)P(x)}{P(1)} \tag{1.14}$$

Comme l'équation (1.14) doit être vraie pour tout b, par dérivation par rapport à b, on doit avoir

$$xP^{'}(bx) = \frac{P^{'}(b)P(x)}{P(1)}.$$

Faisons $b = 1$, on obtient l'équation différentielle,

$$xP^{'}(x) = \frac{P^{'}(1)}{P(1)}P(x),$$

de solution

$$\ln P(x) = \frac{P^{'}(1)}{P(1)} \ln x + cte.$$

Soit

$$P(x) = P(1)x^{-\tau}$$

avec $\tau = -P^{'}(1)/P(1)$.

Par ailleurs, de nombreux phénomènes donnent lieu à des distributions dont le comportement asymptotique (pour x grand) est en loi de puissance.[32] Les exemples suivants sont illustrés par la figure 1.10 :

- La loi de Zipf : la variable X est alors le rang d'un mot dans un texte (en anglais, les mots les plus courants sont "the", "of" ...) et $P(x)$ sa probabilité d'occurrence. Empiriquement, $\tau \simeq 2.2$.
- La distribution des tremblements de terre en fonction de leur magnitude : $\tau \simeq 3.04$.
- La loi de Pareto (1896) donne le nombre d'individus $P(X)$ possédant une certaine "richesse" X. Cette loi semble bien vérifiée dans des situations très diverses pour les revenus les plus élevés. L'exposant τ caractérise alors la violence de l'inégalité sociale. Dans l'exemple de la figure 1.10, $\tau \simeq 2.09$ et la richesse la plus élevée aux USA, 46 10^9\$, est celle d'un certain William H. Gates III. En Inde, pour les années 2002-2004, on trouve $\tau \simeq 1.9$.

.

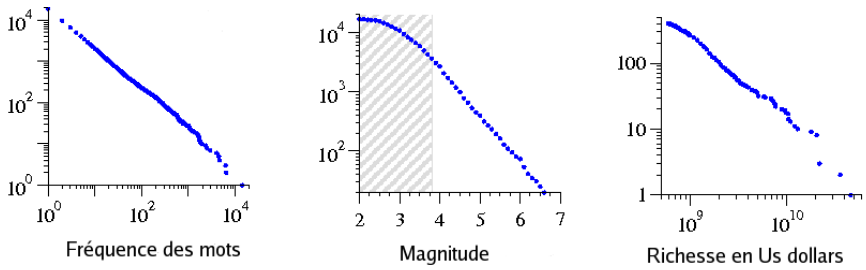

FIG. 1.10 – *Exemples de distributions cumulées ($\int_x^\infty P(x')dx'$, l'exposant vaut donc $\tau - 1$) en loi de puissance (d'après M.E.J. Newman, Contemporary Physics 46, 323-351 (2005)). De gauche à droite : Nombres d'occurrences des mots du roman "Moby Dick" de Hermann Melville ; Magnitudes des tremblements de terre en Californie entre janvier 1910 et mai 1992 ; Revenus en dollars des américains les plus riches en octobre 2003 (voir texte).*

[32] On trouvera divers exemples dans : M.E.J. Newman, "Power laws, Pareto distribution and Zipf's law" , Contemporary Physics 46, 323-351 (2005). Disponible sur http ://fr.arxiv.org/abs/cond-mat/0412004.

1.4 Loi des grands nombres et théorème de la limite centrale

Ces théorèmes jouent un rôle fondamental en statistique, mais ils sont parfois appliqués (à tort) en dehors de leur domaine de validité.

Soit $X_1,\ X_2 \dots\ X_N$ des variables aléatoires mutuellement *indépendantes* ayant la même[33] distribution de probabilité, c'est-à-dire la même moyenne *finie* $\langle X_i \rangle = m$ et la même variance, *pas nécessairement finie*, σ^2. Leur moyenne arithmétique, qui s'écrit

$$M_N = \frac{S_N}{N} = \frac{1}{N}\sum_{i=1}^{N} X_i,$$

est aussi une variable aléatoire. On a évidemment, *en moyenne*, $\langle M_N \rangle = m$ pour tout N : la moyenne de la somme est la somme des moyennes.

Et d'après la relation (1.4), et puisque les variables sont indépendantes :

$$Var(M_N) = \frac{\sigma^2}{N}.$$

1.4.1 La loi des grands nombres

Comme on le voit, lorsque $N \to \infty$, l'écart type tend vers 0 en $1/\sqrt{N}$ et donc

$$\lim_{N\to\infty} M_N = m.$$

Plus formellement, la loi des grands nombres donne la limite en probabilité :

$$\lim_{N\to\infty} Pr\left(|M_N - m| > h\right) = 0 \quad \forall h > 0.$$

Autrement dit, la probabilité que la moyenne M_N s'écarte de m d'une valeur h arbitraire tend vers 0. Ce résultat intuitif a été montré par Khintchine en 1929.

Démontrons ce théorème en deux étapes. Soit X une variable aléatoire de densité de probabilité $P(x)$. On suppose d'abord que $\langle X \rangle = 0$ et que $\langle X^2 \rangle$ existe. On a la suite d'inégalités :

$$Pr(|X| > a) = \int_{|x|>a} P(x)dx < \int_{|x|>a} \left|\frac{x}{a}\right|^2 P(x)dx < \int_{-\infty}^{\infty} \left|\frac{x}{a}\right|^2 P(x)dx = \frac{\langle X^2 \rangle}{a^2}.$$

On suppose maintenant que $\langle X \rangle = m$ non nécessairement nul. On change X en $X - m$ et l'inégalité précédente devient

$$Pr(|X - m| > a) < \frac{\sigma^2}{a^2}.$$

[33] Plus généralement, il n'est pas nécessaire que les N variables aléatoires aient la même moyenne ou la même variance, elles doivent par contre être bornées.

On comprend mieux comment σ caractérise la dispersion de la mesure. Posons maintenant $a = k\sigma$, il vient

$$Pr(|X - m| > k\sigma) < \frac{1}{k^2}.$$

C'est l'inégalité de Bienaymé-Tchebichev.[34] Elle ne nous apprend rien pour $k \leq 1$. On remarquera que cette inégalité a été démontrée en toute généralité, sans connaître la forme de la distribution $P(x)$ (pourvu qu'elle admette un deuxième moment). C'est la majoration la plus grossière possible.

Si on applique maintenant l'inégalité de Bienaymé-Tchebychev, on conclut que[35]

$$Pr\left(|M_N - m| > h\right) < \frac{\sigma^2}{h^2 N},$$

ce qui achève la démonstration.

L'évaluation empirique d'une probabilité est un cas particulier de la loi des grands nombres. En effet, considérons une expérience dont l'issue vaut aléatoirement $X = 1$ avec la probabilité p ou $X = 0$ avec la probabilité $1 - p$. La moyenne arithmétique réalisée sur une série de N observations, $\frac{1}{N}\sum_{i=1}^{N} X_i$, est une estimation de la probabilité p qui est d'autant plus précise que N est grand.

1.4.2 Le théorème de la limite centrale

Ce théorème, démontré par Lindeberg en 1922, va au delà de la connaissance de la valeur moyenne et de la variance de la moyenne arithmétique M_N. Il en donne une loi de probabilité approchée lorsque $N \to \infty$.

[34] Irénée-Jules Bienaymé (1796-1878) mathématicien français et Pafnuty Lvovich Tchebyshev (1821-1894) mathématicien russe.

[35] Si la variance σ^2 n'est pas définie, la démonstration est plus délicate (voir par exemple le livre de Feller, chapitre X).

Soit $X_1,\ X_2 \ldots\ X_N$ des variables aléatoires mutuellement *indépendantes* ayant la même distribution de probabilité,[a] c'est-à-dire la même *moyenne finie* m et la même *variance finie* σ^2. Leur moyenne arithmétique, qui s'écrit

$$M_N = \frac{1}{N}\sum_{i=1}^{N} X_i,$$

est une variable aléatoire distribuée selon une loi de probabilité[b] qui tend vers la loi normale de moyenne m et de variance σ^2/N lorsque $N \to \infty$:

$$P(m_N) \simeq \frac{1}{\sqrt{2\pi}(\frac{\sigma}{\sqrt{N}})} e^{-\frac{(m_N-m)^2}{2(\frac{\sigma}{\sqrt{N}})^2}}, \tag{1.15}$$

[a] Plus généralement, il n'est pas indispensable que ces variables aient la même distribution. Il suffit qu'elles aient la même valeur moyenne et la même variance et qu'elles soient finies.

[b] C'est en fait une densité de probabilité : en passant à la limite $N \to \infty$, M_N devient une variable continue variant de $-\infty$ à $+\infty$.

Insistons : Contrairement à la loi des grands nombres, le théorème de la limite centrale ne s'applique qu'à des variables aléatoires dont la variance est finie.

Nous avons déjà rencontré des applications du théorème de la limite centrale lorsque nous avons étudié le comportement asymptotique de la distribution binomiale et de la loi de Poisson au chapitre 1.3.

On peut vérifier que[†]

$$\lim_{N\to\infty} P(m_N) = \delta(m_N - m).$$

Autrement dit, quand $N \to \infty$, $\sigma/\sqrt{N} \to 0$ et donc la fonction $P(m_N)$ centrée en m devient de plus en plus étroite : son maximum tend vers l'infini, mais son intégrale reste égale à 1. Elle n'a pas de limite comme fonction, mais tend vers la distribution de Dirac, δ, placée en m. La figure 1.11 illustre le théorème de la limite centrale à l'aide de la somme de N variables aléatoires uniformes.

[†] En effet, soit une fonction $f(x)$, il faut montrer que

$$\lim_{N\to\infty} \int f(x)P(x)dx = f(m).$$

On fait le changement de variable $u = (x-m)/(\sqrt{2}\frac{\sigma}{\sqrt{N}})$ et il vient

$$\lim_{N\to\infty} \int f(x)P(x)dx = \frac{1}{\sqrt{\pi}} \int f\left(\frac{\sqrt{2}\sigma}{\sqrt{N}}u + m\right) e^{-u^2} du.$$

Le théorème de la convergence dominée permet de passer à la limite $N \to \infty$, il reste

$$\lim_{N\to\infty} \int f(x)P(x)dx = \frac{1}{\sqrt{\pi}} f(m) \int e^{-u^2} du = f(m).$$

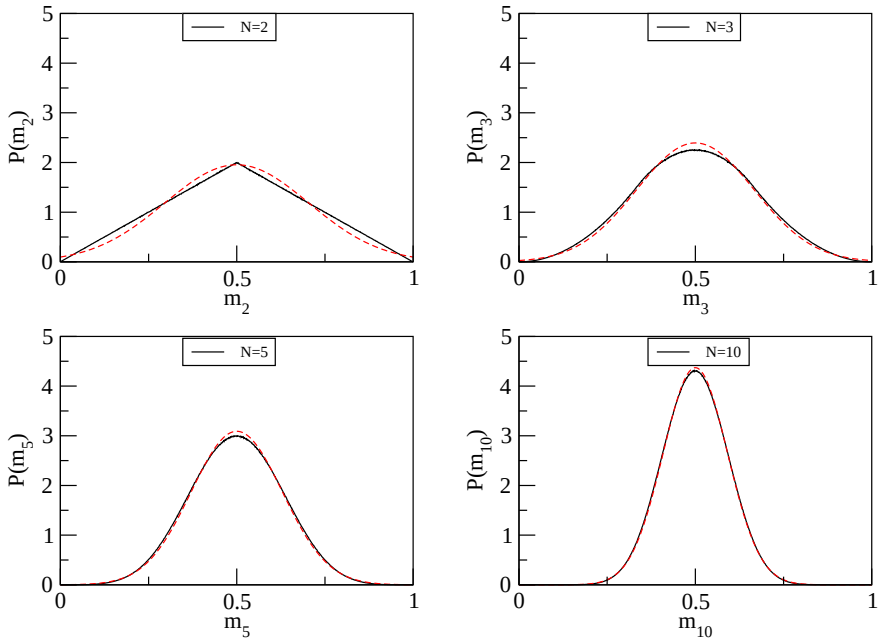

FIG. 1.11 – *Distribution $P(m_N)$ de la somme de $N = 2$ (en haut à gauche), $N = 3$ (en haut à droite), $N = 5$ (en bas à gauche) et $N = 10$ (en bas à droite) variables aléatoires uniformes (distribuées sur $P(x_i) = 1$ pour $0 < x_i < 1$, $Var(X_i) = 1/12$). La courbe en trait rouge interrompu est une gaussienne de même variance que la distribution $P(m_N)$: $Var(M_N) = Var(X_i)/N$.*

Démontrons maintenant le théorème de la limite centrale en utilisant la fonction caractéristique. Puisque M_N est la somme de N variables indépendantes X_i/N, sa fonction caractéristique s'écrit :

$$\phi_{M_N}(u) = \phi^N_{X_i/N}(u), \tag{1.16}$$

où $\phi_{X_i/N}(u)$ est la fonction caractéristique de la variable X_i/N. De la distribution de X_i/N, on ne connaît que ses deux premiers moments : $\langle X_i \rangle = m$ et $\langle X_i^2 \rangle = \sigma^2 + m^2$. En développant $\phi_{X_i/N}(u)$ en série de u on obtient :

$$\begin{aligned} \phi_{X_i/N}(u) &= \langle e^{iu\frac{X_i}{N}} \rangle = 1 + i\frac{u}{N}\langle X_i \rangle - \frac{u^2}{2N^2}\langle X_i^2 \rangle + \mathcal{O}(\frac{u^3}{N^3}) \\ &= 1 + im\frac{u}{N} - \frac{u^2}{2N^2}(\sigma^2 + m^2) + \mathcal{O}(\frac{u^3}{N^3}). \end{aligned}$$

D'après l'équation (1.16), on obtient :

$$\begin{aligned}
\ln \phi_{M_N}(u) &= N \ln \left[1 + im\frac{u}{N} - \frac{u^2}{2N^2}(\sigma^2 + m^2) + \mathcal{O}(\frac{u^3}{N^3}) \right] \\
&= N \left[im\frac{u}{N} - \frac{u^2}{2N^2}(\sigma^2 + m^2) - \frac{1}{2}\big(- \big(\frac{mu}{N}\big)^2\big) + \mathcal{O}(\frac{u^3}{N^3}) \right] \\
&= imu - \frac{u^2\sigma^2}{2N} + \mathcal{O}(\frac{u^3}{N^2}) \\
&= \ln \phi_n(u) + \mathcal{O}(\frac{u^3}{N^2}),
\end{aligned}$$

où $\phi_n(u)$ est la fonction caractéristique de la loi normale de moyenne m et de variance σ^2/N. A la limite $N \to \infty$, la fonction caractéristique de M_N tend vers la fonction caractéristique d'une loi normale, ce qui démontre le théorème de la limite centrale.[36]

En résumé, le théorème de la limite centrale peut s'énoncer de la manière suivante :

> La moyenne arithmétique d'un échantillon d'une population quelconque de variables aléatoires possédant une distribution de probabilité quelconque mais de moyenne et de variance finies, est distribuée suivant une loi normale quand la taille de l'échantillon est suffisamment grande.

Commentaires et conséquences

- Le fait que les distributions de variables de natures très différentes (répétition de la mesure d'une même grandeur physique, taille des individus, variation de la valeur d'une action sur une courte durée, volume de lait produit par une vache...) soient souvent (mais pas toujours !) correctement reproduites par une gaussienne est une donnée empirique. Elle indique seulement que ces variables peuvent résulter d'une somme de variables aléatoires *indépendantes* de même loi. Mais, le mécanisme (complexe) produisant ces variables aléatoires est généralement inconnu.
- Le théorème de la limite centrale ne fait aucune hypothèse sur la forme de la distribution de probabilités des variables X_i, il a donc une portée très générale. En revanche, on ne peut pas l'appliquer lorsque ses deux premiers moments ne sont pas définis.
- Puisque le logarithme d'un produit est la somme des logarithmes des facteurs, le théorème de la limite central implique que le logarithme d'un produit, $\ln Y = \ln \prod_{i=1}^{N} X_i$, de variables aléatoires *indépendantes* et strictement positives tend vers une loi normale. Le produit, Y, d'un *grand nombre* de variables aléatoires indépendantes est donc distribué selon une loi log-normale (voir section 1.3.2).

[36] La rapidité de la convergence vers la gaussienne dépend bien sûr de la distribution des variables X_i. Pour des variables indépendantes et identiquement distribuées, le théorème de Berry-Esséen montre que la convergence est d'autant plus rapide que le rapport $\langle X_i^3 \rangle / \sigma^3$ est petit.

1.4.3 Lois stables

Lorsque la distribution de la somme $S_N = \sum_{i=1}^{N} X_i$ (ou la moyenne arithmétique M_N) a la même dépendance fonctionnelle que celle des variables aléatoires X_i pour N *quelconque* ($N = 2, 3, ...$), on dit que la distribution est *stable*. Notons que la stabilité d'une distribution n'a rien à voir avec le théorème de la limite centrale qui concerne le passage à la limite $N \to \infty$.

Il est facile de montrer à l'aide de la fonction caractéristique et du produit de convolution que les lois suivantes sont stables : normale ($\phi(u) = e^{ium}e^{-u^2\sigma^2/2}$), Poisson ($\phi(u) = e^{\lambda(e^{iu}-1)}$), et Cauchy ($\phi(u)e^{ix_0u-a|u|}$) ... En revanche, comme on l'a vu, la loi uniforme et la loi exponentielle ne sont pas stables.

Ainsi, pour la loi de Cauchy de paramètre a et centrée en x_0, la fonction caractéristique de S_N est

$$\phi_{S_N}(u) = \phi^N(u) = (e^{ix_0u-a|u|})^N.$$

La somme S_N est donc distribuée sur un loi de Cauchy de paramètre aN et centrée en x_0N.

Loi de Lévy

Plus généralement, Lévy et Khintchine ont montré que les fonctions caractéristiques des lois *symétriques* stables ont nécessairement la forme suivante :

$$\phi(u) = e^{imu-a|u|^\alpha},$$

où $0 < \alpha \leq 2$. La loi de Cauchy ($\alpha = 1$) et la gaussienne ($\alpha = 2$) sont donc des cas particuliers de cette loi de Lévy.[37]

La loi de Lévy de moyenne nulle ($m = 0$) et de paramètre $a = 1$ a donc pour expression :

$$\begin{aligned} P_L(x) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \phi(u)e^{-iux}du \\ &= \frac{1}{\pi}\int_0^{+\infty} e^{-|u|^\alpha}\cos(ux)du. \end{aligned}$$

En général, on ne peut pas déterminer la forme explicite de $P_L(x)$. En revanche, on peut montrer que son comportement asymptotique pour $|x|$ grand est en loi de puissance :

$$P_L(x) \sim |x|^{-(1+\alpha)}.$$

Ainsi, les lois de Lévy de paramètre $\alpha < 2$ sont larges.

En finance, les distributions des actifs ne sont pas stables, car leurs paramètres (comme la volatilité) dépendent fortement du temps. En revanche, on peut décrire l'évolution du prix d'un actif financier par un processus infiniment divisible.

[37] Les lois de Lévy asymétriques ont une forme plus compliquée (voir par exemple "An introduction to econophysics" de R.N. Mantegna et H.E. Stanley (Cambridge University press)).

Processus infiniment divisible

Un processus stochastique est infiniment divisible si pour tout entier k il peut être vu comme la somme de k variables aléatoires indépendantes et identiquement distribuées. Autrement dit, la distribution de probabilité $P(x)$ est infiniment divisible si pour tout entier k, il existe une fonction caractéristique $\phi_k(u)$ telle que :

$$\phi(u) = \Big(\phi_k(u)\Big)^k,$$

avec $\phi_k(0) = 1$ et où $\phi(u)$ est la fonction caractéristique de $P(x)$. Un loi stable est don infiniment divisible, mais il en existe d'autres, comme la loi Gamma qui se décompose en une somme de processus exponentiels.

1.5 Théorie de l'information

Développée par Shannon[38] en 1948, la théorie de l'information vise à quantifier la connaissance que l'on a d'un système en estimant *l'information manquante*.

1.5.1 Information et entropie

Soit X une variable aléatoire, la formule de Shannon donne la fonction S_X qui mesure le manque d'information lié à la distribution de probabilité $P(x)$:

$$S_X = -\lambda \langle \ln P(x) \rangle,$$

soit

$$S_X = -\lambda \sum_{x_i} P(x_i) \ln P(x_i) \quad \text{ou} \quad S_X = -\lambda \int P(x) \ln P(x) dx,$$

selon que la variable aléatoire est discrète ou continue et λ est une constante positive qui dépend du contexte. En théorie de l'information, $\lambda = 1/\ln 2$, pour que le manque d'information associé à un ensemble de deux événements équiprobables vaille $S_X = 1$ *bit* ("binary digit"). En physique statistique, on choisit $\lambda = k_b$, la constante de Boltzmann, afin d'identifier l'entropie statistique S_X avec l'entropie thermodynamique.

Dans la suite, on se placera dans le cas où X est discrète. La fonction S_X a été proposée de façon à vérifier les critères suivants :

- $S_X \geq 0$.
- X est certain si et seulement si $S_X = 0$ (il n'y a donc pas de manque d'information).
- Si deux variables aléatoires X et Y sont indépendantes alors

$$S_{X,Y} = -\lambda \sum_{x,y} P(x,y) \ln P(x,y) = S_X + S_Y.$$

[38] Claude Elwood Shannon (1916-2001) mathématicien américain qui fonda la théorie de l'information.

- Si X prend les N valeurs $x_1, x_2, \cdots, x_N$, la distribution de probabilité qui assure S_X maximale (et alors $S_X = \lambda \ln N$) est l'équiprobabilité $P(x_i) = 1/N$, $\forall i$ (voir l'exemple suivant).

En d'autres termes, pour un événement certain, $S_X = 0$ (manque d'information nul) et quand on ne sait rien du système, on suppose l'équiprobabilité des événements : $S_X = \lambda \ln N$ (manque d'information maximal).

1.5.2 Exemples

Pile ou face

Soit une variable aléatoire X qui ne prend que deux valeurs (par exemple 1 et 0) avec les probabilités p et $q = 1 - p$ respectivement. Son entropie s'écrit donc (on prendra $\lambda = 1$)

$$S_X = -p \ln p - (1-p) \ln(1-p).$$

La fonction S_X est maximale (et vaut $\ln 2$) pour $p = 1/2$ (équiprobabilité) et nulle pour $p = 0$ et $p = 1$ (valeurs de X certaines). Lorsque l'on joue à pile ou face, on suppose à priori $p = 1/2$ correspondant à un manque d'information maximal.

Maximisation de l'entropie

On a vu que la distribution qui maximise l'entropie S_X est la loi uniforme. Recherchons à présent la distribution $P(x)$ qui maximise l'entropie pour une variance donnée $Var(X) = \sigma^2$.

Pour simplifier, plaçons nous dans le cas d'une distribution centrée. A la contrainte de la normalisation ($\int dx' P(x') = 1$), s'ajoute donc

$$Var(X) = \langle X^2 \rangle = \int dx' x'^2 P(x') = \sigma^2.$$

En utilisant la méthode du multiplicateur de Lagrange, (voir annexe A.5), on doit donc chercher la distribution $P(x)$ qui maximise (en prenant $\lambda = 1$) :

$$-\int dx' P(x') \ln P(x') - \mu_1 \Big(\int dx' P(x') - 1 \Big) - \mu_2 \Big(\int dx' x'^2 P(x') - \sigma^2 \Big),$$

où μ_1 et μ_2 sont les deux multiplicateurs de Lagrange fixés a posteriori par les deux contraintes. En dérivant par rapport à $P(x)$, on trouve :

$$-\ln P(x) - 1 - \mu_1 - \mu_2 x^2 = 0,$$

soit

$$P(x) = A e^{-B x^2},$$

où A et B sont deux constantes à déterminer avec les deux contraintes. On trouve facilement $A = 1/\sqrt{2\pi\sigma^2}$ et $B = 1/2\sigma^2$. La distribution qui maximise

l'entropie (le manque d'information) pour une variance donnée σ^2 est donc la loi normale. On en déduit l'entropie de la loi normale :

$$S_X = -\int dx' P(x') \ln P(x') = \ln(\sigma\sqrt{2\pi e})$$

On retrouve l'idée du théorème de la limite centrale : en convoluant des variables aléatoires indépendantes, on perd de l'information.

On montrera de la même façon que la loi exponentielle est la distribution dont la moyenne est fixée ($\langle X \rangle = m$) qui maximise l'entropie. On trouve dans ce cas une entropie $S_X = \ln(m\, e)$.