Mes Notes de Lecture

Introduction à la Probabilité

Lou Brunet

29 octobre 2025

Table des matières

1 Probabilités et Dénombrement

1.1 Concepts fondamentaux

Avant de pouvoir calculer des probabilités, il est essentiel d'établir un vocabulaire commun pour décrire les expériences aléatoires.

Intuition : Nécessité d'un Cadre Formel
Avant de calculer des probabilités, il est crucial de définir les règles du jeu :
Qu'est-ce qui peut arriver ?
On définit l'ensemble de tous les résultats possibles de l'expérience.
À quoi s'intéresse-t-on ?
On identifie les sous-ensembles de résultats spécifiques qui nous intéressent.
Ces deux idées nous conduisent aux notions d'Univers et d'Événement, qui sont les piliers de toute théorie des probabilités.

Cette intuition se traduit formellement par deux définitions clés :

Définition : Concepts Fondamentaux

Univers (ou Espace Échantillon), S :
L'ensemble de tous les résultats possibles d'une expérience aléatoire.
Événement, A :
Un sous-ensemble de l'univers ($A \subseteq S$). C'est un ensemble de résultats auxquels on s'intéresse.

Un exemple simple permet de solidifier ces concepts :

Exemple : Univers et Événement
Pour l'expérience du "lancer d'un dé à six faces" :
L'**univers** est $S = \{1, 2, 3, 4, 5, 6\}$.
"Obtenir un nombre impair" est un événement, représenté par le sous-ensemble $A = \{1, 3, 5\}$.

1.2 Définition Naïve de la Probabilité

Pour de nombreuses expériences simples, comme lancer un dé non pipé, chaque résultat possible est "équiprobable". Cette hypothèse est la base de la première définition formelle de la probabilité.

Définition : Probabilité Naïve
Pour une expérience où chaque issue dans un espace échantillon fini S est équiprobable, la probabilité d'un événement A est le rapport du nombre d'issues favorables à A sur le nombre total d'issues :

$$P(A) = \frac{\text{Nombre d'issues favorables}}{\text{Nombre total d'issues}} = \frac{|A|}{|S|}$$

Appliquons cette formule à quelques cas classiques :

Exemple : Applications de la définition naïve

1. **Lancer une pièce équilibrée :** L'espace échantillon est $S = \{\text{Pile, Face}\}$, donc $|S| = 2$. Si l'événement A est "obtenir Pile", alors $A = \{\text{Pile}\}$ et $|A| = 1$. La probabilité est $P(A) = \frac{1}{2}$.
2. **Lancer un dé à six faces non pipé :** L'espace échantillon est $S = \{1, 2, 3, 4, 5, 6\}$, donc $|S| = 6$. Si l'événement B est "obtenir un nombre pair", alors $B = \{2, 4, 6\}$ et $|B| = 3$. La probabilité est $P(B) = \frac{3}{6} = \frac{1}{2}$.
3. **Tirer une carte d'un jeu de 52 cartes :** L'espace échantillon S contient 52 cartes, donc $|S| = 52$. Si l'événement C est "tirer un Roi", il y a 4 Rois dans le jeu, donc $|C| = 4$. La probabilité est $P(C) = \frac{4}{52} = \frac{1}{13}$.

1.3 Permutations (Arrangements)

Le dénombrement, qui est l'art de compter les tailles $|A|$ et $|S|$, est fondamental pour appliquer la définition naïve. Le premier outil que nous verrons est la permutation, qui compte les arrangements **ordonnés**.

Définition : Permutation de k objets parmi n

Le nombre de façons d'arranger k objets choisis parmi n objets distincts (où l'ordre compte et il n'y a pas de répétition) est noté $P(n,k)$ ou A_n^k et est défini par :

$$P(n,k) = \frac{n!}{(n-k)!}$$

où $n!$ est la factorielle de n, et par convention $0! = 1$.

Cette formule peut sembler abstraite, mais elle provient d'un raisonnement logique simple par "cases" :

Intuition : Permutations de k parmi n

Pour placer k objets dans un ordre spécifique en les choisissant parmi n objets disponibles, on a n choix pour la première position, $(n-1)$ choix pour la deuxième, ..., et $(n-k+1)$ choix pour la k-ième position. Cela donne $n \times (n-1) \times \cdots \times (n-k+1)$ arrangements. Ce produit contient k termes. Il est égal à $\frac{n!}{(n-k)!}$, car cela revient à diviser la suite complète $n!$ par les facteurs non utilisés $(n-k) \times (n-k-1) \times \cdots \times 1$.

Voyons une application classique de ce principe :

Exemple : Permutations de k parmi n

Podium d'une course : Une course réunit 8 coureurs. Combien y a-t-il de podiums (1er, 2e, 3e) possibles ?
On cherche le nombre de façons d'ordonner 3 coureurs parmi 8 : $P(8,3)$.

$$P(8,3) = \frac{8!}{(8-3)!} = \frac{8!}{5!} = 8 \times 7 \times 6 = 336$$

Il y a 336 podiums possibles.

1.4 Le Coefficient Binomial

Que se passe-t-il si l'ordre ne compte pas ? Au lieu de compter des podiums, nous voulons compter des comités. C'est le rôle du coefficient binomial.

Théorème : Formule du Coefficient Binomial

Le nombre de façons de choisir k objets parmi un ensemble de n objets distincts (sans remise et sans ordre) est donné par le coefficient binomial :

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

La preuve de cette formule repose sur un argument combinatoire élégant : nous allons compter la même chose (les permutations) de deux façons différentes.

Preuve

Considérons le nombre de permutations de k objets parmi n, noté $P(n,k)$.

1. **Méthode 1 :** Par définition (vue ci-dessus), nous savons que $P(n,k) = \frac{n!}{(n-k)!}$.
2. **Méthode 2 :** Nous pouvons construire une telle permutation en deux étapes successives :
   - D'abord, **choisir un sous-ensemble** de k objets parmi n (l'ordre ne compte pas). C'est le nombre que nous cherchons, notons-le $\binom{n}{k}$.
   - Ensuite, **ordonner** ces k objets choisis. Il y a $k!$ façons de les arranger.

   Le nombre total de permutations est donc le produit de ces étapes : $P(n,k) = \binom{n}{k} \times k!$.

En égalisant les deux méthodes, on obtient :

$$\binom{n}{k} \cdot k! = \frac{n!}{(n-k)!}$$

En divisant par $k!$, on trouve bien la formule :

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

L'intuition visuelle derrière cette preuve est de voir comment chaque "choix" (une colonne du tableau) génère $k!$ "ordres" (les lignes de cette colonne).

Intuition

Pour rendre cela concret, voici le cas $\binom{5}{3}$. Il y a 10 sous-ensembles de 3 éléments parmi $\{a,b,c,d,e\}$. Chacun donne lieu à $3! = 6$ permutations. Le tableau ci-dessous montre **toutes les 60 permutations**, regroupées par sous-ensemble :

$\{a,b,c\}$	$\{a,b,d\}$	$\{a,b,e\}$	$\{a,c,d\}$	$\{a,c,e\}$	$\{a,d,e\}$	$\{b,c,d\}$	$\{b,c,e\}$	$\{b,d,e\}$	$\{c,d,e\}$
abc	*abd*	*abe*	*acd*	*ace*	*ade*	*bcd*	*bce*	*bde*	*cde*
acb	*adb*	*aeb*	*adc*	*aec*	*aed*	*bdc*	*bec*	*bed*	*ced*
bac	*bad*	*bae*	*cad*	*cae*	*dae*	*cbd*	*ceb*	*dbe*	*dce*
bca	*bda*	*bea*	*cda*	*cea*	*dea*	*cdb*	*ceb*	*deb*	*dec*
cab	*dab*	*eab*	*dac*	*eac*	*ead*	*dbc*	*ebc*	*edb*	*ecd*
cba	*dba*	*eba*	*dca*	*eca*	*eda*	*dcb*	*ebc*	*edb*	*edc*

Chaque colonne correspond à **un seul et même choix non ordonné** (par exemple $\{a,b,c\}$), mais à 6 listes différentes selon l'ordre. Ainsi, pour obtenir le nombre de *choix non ordonnés*, on divise le nombre total de listes (60) par le nombre d'ordres par groupe (6) :

$$\binom{5}{3} = \frac{60}{6} = 10.$$

L'application la plus directe est le tirage d'un groupe où l'ordre n'importe pas :

Exemple : Utilisation du Coefficient Binomial

Comité d'étudiants : De combien de manières peut-on former un comité de 3 étudiants à partir d'une classe de 10 ? L'ordre ne compte pas.

$$\binom{10}{3} = \frac{10!}{3!(10-3)!} = \frac{10 \times 9 \times 8}{3 \times 2 \times 1} = 120 \text{ comités possibles.}$$

1.5 Identité de Vandermonde

Les coefficients binomiaux obéissent à de nombreuses identités. L'identité de Vandermonde est l'une des plus utiles, car elle montre comment décomposer un problème de comptage complexe en sous-problèmes.

Théorème : Identité de Vandermonde

Cette identité offre une relation remarquable entre les coefficients binomiaux. Pour des entiers non négatifs m, n et k, on a :

$$\binom{m+n}{k} = \sum_{j=0}^{k} \binom{m}{j}\binom{n}{k-j}$$

La preuve la plus intuitive est une "preuve par l'histoire" (proof by story), qui consiste à trouver un scénario de dénombrement que les deux côtés de l'équation résolvent.

Preuve : Preuve combinatoire

Imaginons un groupe composé de m hommes et n femmes. Nous souhaitons former un comité de k personnes. Nous allons compter le nombre de comités possibles de deux façons.

Côté gauche : $\binom{m+n}{k}$ Le groupe total contient $m+n$ personnes. Le nombre de façons de choisir un comité de k personnes parmi ce total est, par définition, $\binom{m+n}{k}$.

Côté droit : $\sum_{j=0}^{k}\binom{m}{j}\binom{n}{k-j}$ Nous pouvons compter le même nombre en conditionnant sur le nombre d'hommes (noté j) dans le comité. Un comité de k personnes doit contenir j hommes ET $k-j$ femmes, où j peut aller de 0 à k.

- Pour $j=0$: Choisir 0 homme ($\binom{m}{0}$) ET k femmes ($\binom{n}{k}$).
- Pour $j=1$: Choisir 1 homme ($\binom{m}{1}$) ET $k-1$ femmes ($\binom{n}{k-1}$).
- ...
- Pour $j=k$: Choisir k hommes ($\binom{m}{k}$) ET 0 femme ($\binom{n}{0}$).

Puisque ces cas (0 homme, 1 homme, etc.) sont mutuellement exclusifs, le nombre total de comités est la somme de toutes ces possibilités :

$$\sum_{j=0}^{k} \binom{m}{j}\binom{n}{k-j}$$

Puisque les deux côtés comptent exactement la même chose (le nombre total de comités), ils doivent être égaux.

Vérifions cette identité avec un exemple numérique concret, en reprenant l'analogie du comité :

Exemple : Application de l'Identité de Vandermonde

On veut former un comité de 3 personnes ($k=3$) à partir d'un groupe de 5 hommes ($m=5$) et 4 femmes ($n=4$).

Méthode directe (côté gauche) : On choisit 3 personnes parmi les $5+4=9$ au total.

$$\binom{9}{3} = \frac{9\times 8\times 7}{3\times 2\times 1} = 84$$

Méthode par cas (côté droit) : La somme est $\binom{5}{0}\binom{4}{3} + \binom{5}{1}\binom{4}{2} + \binom{5}{2}\binom{4}{1} + \binom{5}{3}\binom{4}{0} = 84$. Les deux méthodes donnent bien le même résultat.

1.6 Bose-Einstein (Étoiles et Bâtons)

Jusqu'à présent, nous avons supposé un "tirage sans remise". La statistique de Bose-Einstein, ou plus visuellement la méthode des "étoiles et bâtons", s'attaque au problème du **tirage avec remise** où l'ordre ne compte pas.

Théorème : Combinaisons avec répétition

Le nombre de façons de distribuer k objets indiscernables dans n boîtes discernables (ou de choisir k objets parmi n avec remise, où l'ordre ne compte pas) est donné par la formule :

$$\binom{n+k-1}{k} = \binom{n+k-1}{n-1}$$

La preuve de cette formule est l'un des résultats les plus élégants du dénombrement. L'astuce consiste à transformer le problème de distribution en un problème d'arrangement de symboles.

Preuve : Par les "Étoiles et Bâtons"

Nous cherchons à distribuer k objets indiscernables ($\star$) dans n boîtes discernables. Nous pouvons représenter n'importe quelle distribution comme une séquence de symboles. Nous avons besoin de k étoiles (les objets) et de $n-1$ bâtons ($|$) pour servir de séparateurs entre les n boîtes.
Par exemple, pour distribuer $k = 7$ étoiles dans $n = 4$ boîtes, la séquence :

$$\star\star\star \,|\, \star \,||\, \star\star\star$$

correspond à : 3 étoiles dans la boîte 1, 1 étoile dans la boîte 2, 0 étoile dans la boîte 3 (l'espace entre deux bâtons), et 3 étoiles dans la boîte 4.
Chaque arrangement unique de ces symboles correspond à une distribution unique. Le problème revient donc à trouver le nombre de façons d'arranger ces k étoiles et ces $n-1$ bâtons.
Nous avons un total de $n+k-1$ positions à remplir. Le nombre de façons de le faire est simplement le nombre de manières de choisir les k positions pour les étoiles (les autres positions étant automatiquement remplies par des bâtons). C'est exactement :

$$\binom{n+k-1}{k}$$

(Ce qui est aussi égal à $\binom{n+k-1}{n-1}$, le nombre de façons de choisir les positions des $n-1$ bâtons).

C'est la méthode parfaite pour tout problème de distribution d'objets identiques :

Exemple : Distribution de biens identiques

De combien de manières peut-on distribuer 10 croissants identiques à 4 enfants ?
Ici, $k = 10$ (les croissants, objets indiscernables) et $n = 4$ (les enfants, boîtes discernables). Le nombre de distributions possibles est :

$$\binom{4+10-1}{10} = \binom{13}{10} = \binom{13}{3} = \frac{13 \times 12 \times 11}{3 \times 2 \times 1} = 13 \times 2 \times 11 = 286$$

Il y a 286 façons de distribuer les croissants.

1.7 Principe d'Inclusion-Exclusion

Comment compter le nombre d'éléments dans l'union de plusieurs ensembles ? Si on additionne simplement leurs tailles, on compte les intersections plusieurs fois. Le principe d'inclusion-exclusion corrige systématiquement ce sur-comptage.

Théorème : Principe d'Inclusion-Exclusion pour 3 ensembles

Pour trois ensembles finis A, B et C, le nombre d'éléments dans leur union est donné par :

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$$

La preuve pour 3 ensembles se fait en appliquant la formule pour 2 ensembles de manière répétée.

Preuve

Nous utilisons la formule pour deux ensembles, $|X \cup Y| = |X| + |Y| - |X \cap Y|$, de manière imbriquée. Posons $X = A \cup B$ et $Y = C$.

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

Nous devons maintenant développer les deux termes compliqués :

1. $|A \cup B| = |A| + |B| - |A \cap B|$
2. Par distributivité de l'intersection sur l'union, $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$.

Appliquons la formule pour 2 ensembles à ce deuxième terme :

$$|(A \cap C) \cup (B \cap C)| = |A \cap C| + |B \cap C| - |(A \cap C) \cap (B \cap C)|$$

Ce qui se simplifie en $|A \cap C| + |B \cap C| - |A \cap B \cap C|$.
Finalement, en substituant tout dans l'équation de départ :

$$\begin{aligned} |A \cup B \cup C| = &\underbrace{(|A| + |B| - |A \cap B|)}_{|A \cup B|} + |C| \\ &- \underbrace{(|A \cap C| + |B \cap C| - |A \cap B \cap C|)}_{|(A \cup B) \cap C|} \end{aligned}$$

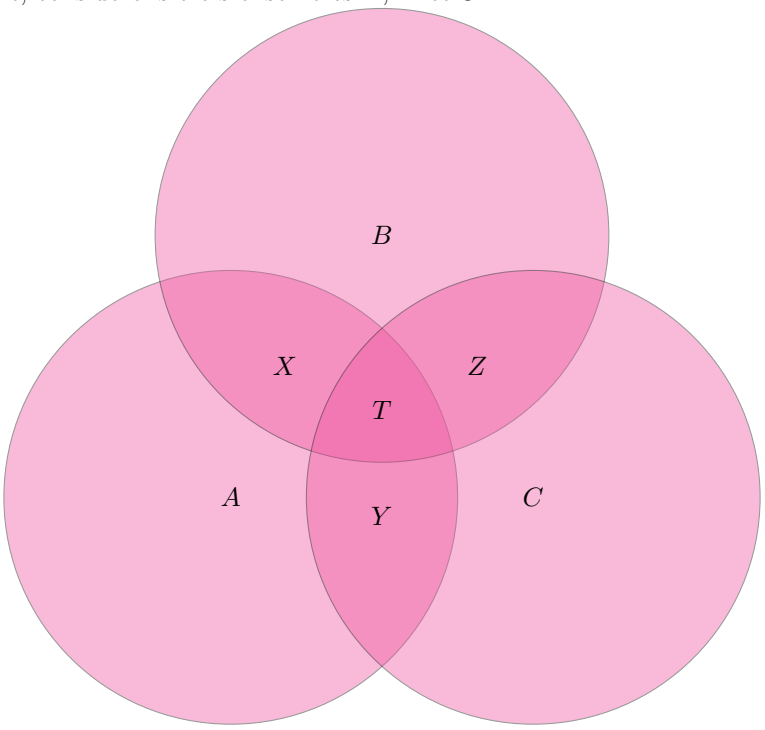
En réarrangeant les termes, on obtient la formule voulue :

$$|A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$$

La formule devient évidente lorsque l'on utilise un diagramme de Venn pour visualiser le sur-comptage et sa correction.

Intuition : Visualisation avec 3 ensembles

Le principe d'inclusion-exclusion permet de compter le nombre d'éléments dans une union d'ensembles sans double-comptage. Pour comprendre intuitivement pourquoi on ajoute et soustrait alternativement, considérons trois ensembles A, B et C :



Le problème : Si on additionne simplement $|A|+|B|+|C|$, on compte certaines zones plusieurs fois :

- Les intersections deux à deux (X, Y, Z) sont comptées **deux fois**
- L'intersection triple (T) est comptée **trois fois**

La solution : On corrige en soustrayant les intersections deux à deux, mais alors l'intersection triple est comptée :

- $+3$ fois dans la somme initiale
- -3 fois dans la soustraction des intersections deux à deux (car elle appartient à chacune)
- Donc 0 fois au total ! Il faut la rajouter.

D'où la formule : $|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$

Ce que nous avons fait visuellement pour 3 ensembles peut être généralisé par récurrence à n ensembles. La formule générale suit le même principe d'alternance des signes :

Théorème : Principe d'Inclusion-Exclusion généralisé

Pour n ensembles finis $A_1, A_2, \ldots, A_n$, on a :

$$\begin{aligned} |A_1 \cup A_2 \cup \cdots \cup A_n| = &\sum_{i=1}^{n} |A_i| \\ &- \sum_{1 \le i < j \le n} |A_i \cap A_j| \\ &+ \sum_{1 \le i < j < k \le n} |A_i \cap A_j \cap A_k| \\ &- \cdots \\ &+ (-1)^{n+1} |A_1 \cap A_2 \cap \cdots \cap A_n| \end{aligned}$$

Ce qui s'écrit plus compactement :

$$\left| \bigcup_{i=1}^{n} A_i \right| = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} |A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}|$$

La preuve formelle que cette formule gigantesque fonctionne est fascinante. Il suffit de montrer que n'importe quel élément x de l'union, peu importe à combien d'ensembles il appartient, est compté **exactement une fois** au final.

Preuve : Preuve par comptage d'un élément

Considérons un élément x qui appartient à exactement k ensembles parmi les n ensembles $A_1, \ldots, A_n$ (où $k \ge 1$). Nous devons montrer que x est compté exactement 1 fois par la formule.

Analysons combien de fois x est compté dans chaque somme de la formule :

- **Première somme** ($\sum |A_i|$) : x est dans k ensembles, donc il est ajouté k fois. Le nombre de fois est $\binom{k}{1}$.
- **Deuxième somme** ($-\sum |A_i \cap A_j|$) : x est compté (et soustrait) pour chaque *paire* d'ensembles auxquels il appartient. Comme il appartient à k ensembles, il y a $\binom{k}{2}$ telles paires.
- **Troisième somme** ($+\sum |A_i \cap A_j \cap A_k|$) : x est ajouté pour chaque *triplet* d'ensembles auxquels il appartient. Il y en a $\binom{k}{3}$.
- **Et ainsi de suite...**

Au total, l'élément x est compté :

$$\text{Total} = \binom{k}{1} - \binom{k}{2} + \binom{k}{3} - \cdots + (-1)^{k-1} \binom{k}{k} \text{ fois.}$$

Pour évaluer cette somme, rappelons l'identité fondamentale du binôme de Newton :

$$(1+x)^k = \sum_{j=0}^{k} \binom{k}{j} x^j = \binom{k}{0} + \binom{k}{1}x + \binom{k}{2}x^2 + \cdots$$

Si nous posons $x = -1$, nous obtenons :

$$(1-1)^k = 0 = \binom{k}{0} - \binom{k}{1} + \binom{k}{2} - \binom{k}{3} + \cdots + (-1)^k \binom{k}{k}$$

Sachant que $\binom{k}{0} = 1$, on a :

$$0 = 1 - \left(\binom{k}{1} - \binom{k}{2} + \binom{k}{3} - \cdots + (-1)^{k-1} \binom{k}{k} \right)$$

En réarrangeant, on trouve :

$$1 = \binom{k}{1} - \binom{k}{2} + \binom{k}{3} - \cdots + (-1)^{k-1} \binom{k}{k}$$

Cela prouve que n'importe quel élément de l'union est compté exactement une fois.

Ce principe est très utile en probabilité, car il permet de calculer $P(A \cup B \cup \dots)$ en se basant sur les probabilités des intersections, qui sont souvent plus faciles à trouver.

Exemple : Application probabiliste

On lance trois dés équilibrés. Quelle est la probabilité d'obtenir au moins un 6 ?
Solution avec inclusion-exclusion :
Soit $A =$ "le premier dé montre 6", $B =$ "le deuxième dé montre 6", $C =$ "le troisième dé montre 6".
On veut $P(A \cup B \cup C)$.

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &\quad - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &\quad + P(A \cap B \cap C) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} - \frac{1}{36} - \frac{1}{36} - \frac{1}{36} + \frac{1}{216} \\ &= \frac{3}{6} - \frac{3}{36} + \frac{1}{216} = \frac{1}{2} - \frac{1}{12} + \frac{1}{216} \\ &= \frac{108 - 18 + 1}{216} = \frac{91}{216} \approx 0.421 \end{aligned}$$

Vérification par la méthode complémentaire :
La probabilité de n'obtenir aucun 6 est $\left(\frac{5}{6}\right)^3 = \frac{125}{216}$, donc la probabilité d'au moins un 6 est $1 - \frac{125}{216} = \frac{91}{216}$.

2 Probabilité conditionnelle

Intuition : Question Fondamentale

La probabilité conditionnelle est le concept qui répond à la question fondamentale : comment devons-nous mettre à jour nos croyances à la lumière des nouvelles informations que nous observons ?

Ce concept de "mise à jour des croyances" est le cœur de la statistique moderne. Il s'agit de quantifier comment une nouvelle information B affecte la probabilité d'un événement A.

2.1 Définition de la Probabilité Conditionnelle

Commençons par la définition formelle.

Définition : Probabilité Conditionnelle

Si A et B sont deux événements avec $P(B) > 0$, alors la probabilité conditionnelle de A sachant B, notée $P(A|B)$, est définie comme :

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

Cette formule n'est pas sortie de nulle part. Elle représente une "réduction de l'univers" :

Intuition

Imaginez que l'ensemble de tous les résultats possibles est un grand terrain. Savoir que l'événement B s'est produit, c'est comme si on vous disait que le résultat se trouve dans une zone spécifique de ce terrain. La probabilité conditionnelle $P(A|B)$ ne s'intéresse plus au terrain entier, mais seulement à la proportion de la zone B qui est également occupée par A. On "zoome" sur le monde où B est vrai, et on recalcule les probabilités dans ce nouveau monde plus petit.

2.2 Règle du Produit (Intersection de deux événements)

En réarrangeant simplement les termes de la définition, nous obtenons une règle fondamentale pour calculer la probabilité que deux événements se produisent *ensemble*.

Théorème : Probabilité de l'intersection de deux événements

Pour tous événements A et B avec des probabilités positives, nous avons :

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

Cela découle directement de la définition de la probabilité conditionnelle.

La preuve est une simple réorganisation algébrique :

Preuve

La preuve est une simple réorganisation algébrique de la définition de la probabilité conditionnelle. Par définition, nous avons :

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

En multipliant les deux côtés par $P(B)$, on obtient :

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

De même, en partant de $P(B|A) = \frac{P(B \cap A)}{P(A)}$, on obtient :

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

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

Cette formule exprime mathématiquement l'idée séquentielle suivante :

Intuition
Pour que deux événements se produisent, le premier doit se produire, PUIS le second doit se produire, sachant que le premier a eu lieu.

Cette règle est particulièrement utile pour les tirages sans remise, où la probabilité du second événement dépend du résultat du premier.

Exemple
Quelle est la probabilité de tirer deux As d'un jeu de 52 cartes sans remise ? Soit A l'événement "le premier tirage est un As", avec $P(A) = \frac{4}{52}$. Soit B l'événement "le deuxième tirage est un As". Nous cherchons $P(A \cap B)$, que l'on calcule avec la formule $P(A \cap B) = P(A) \times P(B|A)$. La probabilité $P(B|A)$ correspond à tirer un As sachant que la première carte était un As. Il reste alors 51 cartes, dont 3 As. Donc, $P(B|A) = \frac{3}{51}$. Finalement, la probabilité de l'intersection est $P(A \cap B) = \frac{4}{52} \times \frac{3}{51} = \frac{12}{2652} \approx 0.0045$.

2.3 Règle de la Chaîne (Intersection de n événements)

On peut logiquement étendre cette règle de deux à n événements.

Théorème : Probabilité de l'intersection de n événements
Pour tous événements $A_1, \ldots, A_n$ avec $P(A_1 \cap A_2 \cap \cdots \cap A_{n-1}) > 0$, nous avons :

$$P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap \cdots \cap A_{n-1})$$

La preuve se fait par une simple récurrence :

Preuve : Preuve par récurrence
Nous pouvons prouver cela par une application répétée de la règle du produit pour deux événements.
Cas de base (n=2) : $P(A_1 \cap A_2) = P(A_1)P(A_2|A_1)$. C'est le théorème précédent.
Étape (n=3) : Traitons $(A_1 \cap A_2)$ comme un seul événement :

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P((A_1 \cap A_2) \cap A_3) \\ &= P(A_1 \cap A_2) \times P(A_3|A_1 \cap A_2) \\ &= (P(A_1)P(A_2|A_1)) \times P(A_3|A_1 \cap A_2) \end{aligned}$$

Généralisation : En continuant ce processus, on voit que pour ajouter A_n, on multiplie par la probabilité de A_n conditionnée par l'intersection de tous les événements précédents $(A_1 \cap \cdots \cap A_{n-1})$.

Cette "règle de la chaîne" (chain rule) est cruciale pour les processus stochastiques :

Intuition
Pour qu'une séquence d'événements se produise, chaque événement doit se réaliser tour à tour, en tenant compte de tous les événements précédents qui se sont déjà produits.

Reprenons l'exemple des cartes, mais en continuant le tirage :

Exemple
On tire 3 cartes sans remise. Quelle est la probabilité d'obtenir la séquence Roi, Dame, Valet ? La probabilité de tirer un Roi en premier (A_1) est $P(A_1) = \frac{4}{52}$. Ensuite, la probabilité de tirer une Dame (A_2) sachant qu'un Roi a été tiré est $P(A_2|A_1) = \frac{4}{51}$. Enfin, la probabilité de tirer un Valet (A_3) sachant qu'un Roi et une Dame ont été tirés est $P(A_3|A_1 \cap A_2) = \frac{4}{50}$. La probabilité totale de la séquence est donc le produit de ces probabilités : $P(A_1 \cap A_2 \cap A_3) = \frac{4}{52} \times \frac{4}{51} \times \frac{4}{50} \approx 0.00048$.

2.4 Règle de Bayes

La règle du produit est aussi la pierre angulaire de la formule la plus célèbre des probabilités conditionnelles, qui nous permet d'inverser la condition.

Théorème : Règle de Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

La preuve est élégante car elle utilise simplement la symétrie de l'intersection :

Preuve

La preuve découle de l'égalité de la règle du produit. Nous savons que :

1. $P(A \cap B) = P(A|B)P(B)$
2. $P(A \cap B) = P(B|A)P(A)$

En égalisant ces deux expressions, on a :

$$P(A|B)P(B) = P(B|A)P(A)$$

En supposant $P(B) > 0$ et en divisant par $P(B)$, on obtient :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

L'importance de cette formule ne peut être sous-estimée :

Intuition

La règle de Bayes est la formule pour "inverser" une probabilité conditionnelle. Souvent, il est facile de connaître la probabilité d'un effet étant donné une cause (P(symptôme|maladie)), mais ce qui nous intéresse vraiment, c'est la probabilité de la cause étant donné l'effet observé (P(maladie|symptôme)). La règle de Bayes nous permet de faire ce retournement en utilisant notre connaissance initiale de la probabilité de la cause (P(maladie)). C'est le fondement mathématique de la mise à jour de nos croyances.

2.5 Formule des Probabilités Totales

Le dénominateur $P(B)$ dans la règle de Bayes est souvent inconnu. Pour le trouver, nous avons besoin d'un autre outil puissant.

Théorème : Formule des probabilités totales

Soit $A_1, \ldots, A_n$ une partition de l'espace échantillon S (c'est-à-dire que les A_i sont des événements disjoints et leur union est S), avec $P(A_i) > 0$ pour tout i. Alors pour tout événement B :

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

La démonstration repose sur la décomposition de l'événement B sur la partition A_i.

Preuve : Démonstration de la formule des probabilités totales

Puisque les A_i forment une partition de S, on peut décomposer B comme :

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_n)$$

Comme les A_i sont disjoints, les événements $(B \cap A_i)$ le sont aussi. On peut donc sommer leurs probabilités :

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_n)$$

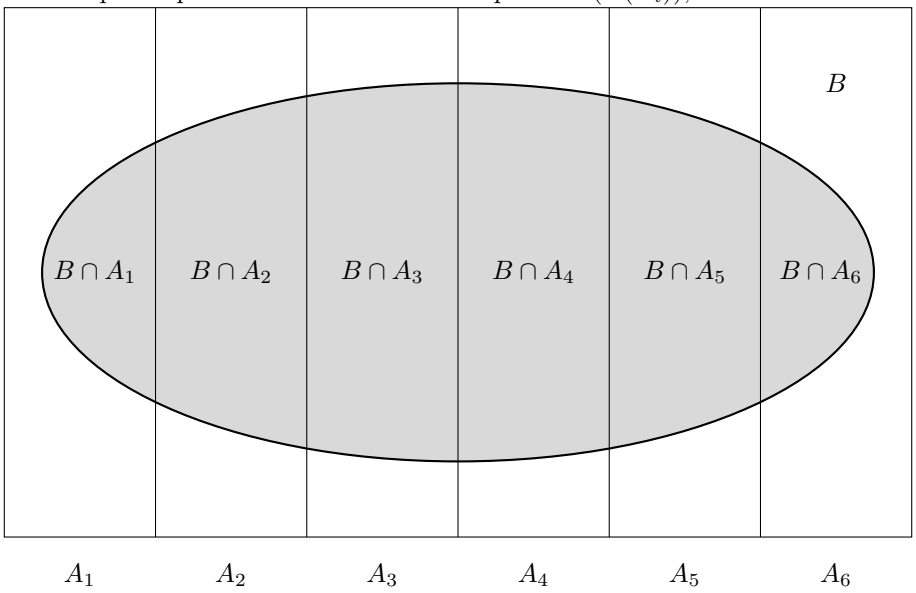
En appliquant le théorème de l'intersection des probabilités à chaque terme, on obtient :

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_n) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Visuellement, cette formule consiste à "découper" l'événement B et à additionner les morceaux :

Intuition

C'est une stratégie de "diviser pour régner". Pour calculer la probabilité totale d'un événement B, on peut décomposer le monde en plusieurs scénarios mutuellement exclusifs (la partition A_i). On calcule ensuite la probabilité de B dans chacun de ces scénarios ($P(B|A_i)$), on pondère chaque résultat par la probabilité du scénario en question ($P(A_i)$), et on additionne le tout.



L'exemple de l'usine est un cas d'école pour cette formule :

Exemple

Une usine possède trois machines, M1, M2, et M3, qui produisent respectivement 50%, 30% et 20% des articles. Leurs taux de production défectueuse sont de 4%, 2% et 5%. Quelle est la probabilité qu'un article choisi au hasard soit défectueux ? Soit D l'événement "l'article est défectueux". Les machines forment une partition avec $P(M1) = 0.5$, $P(M2) = 0.3$, et $P(M3) = 0.2$. Les probabilités conditionnelles de défaut sont $P(D|M1) = 0.04$, $P(D|M2) = 0.02$, et $P(D|M3) = 0.05$. En appliquant la formule, on obtient : $P(D) = P(D|M1)P(M1) + P(D|M2)P(M2) + P(D|M3)P(M3) = (0.04 \times 0.5) + (0.02 \times 0.3) + (0.05 \times 0.2) = 0.02 + 0.006 + 0.01 = 0.036$. La probabilité qu'un article soit défectueux est de 3.6%.

Maintenant, nous pouvons combiner la Règle de Bayes et la Formule des Probabilités Totales pour résoudre des problèmes complexes, comme celui du dépistage médical.

Exemple : Application Combinée : Bayes et Probabilités Totales

Une maladie touche 1% de la population ($P(M) = 0.01$). Un test de dépistage est fiable à 95% : il est positif pour 95% des malades ($P(T|M) = 0.95$) et négatif pour 95% des non-malades, ce qui implique un taux de faux positifs de $P(T|\neg M) = 0.05$. Une personne est testée positive. Quelle est la probabilité qu'elle soit réellement malade, $P(M|T)$?
On cherche $P(M|T) = \frac{P(T|M)P(M)}{P(T)}$.
D'abord, on calcule $P(T)$ avec la formule des probabilités totales (la partition est $\{M, \neg M\}$) : $P(T) = P(T|M)P(M) + P(T|\neg M)P(\neg M) = (0.95 \times 0.01) + (0.05 \times 0.99) = 0.0095 + 0.0495 = 0.059$.
Ensuite, on applique la règle de Bayes : $P(M|T) = \frac{0.95 \times 0.01}{0.059} \approx 0.161$. Malgré un test positif, il n'y a que 16.1% de chance que la personne soit malade.

2.6 Règle de Bayes avec Conditionnement Additionnel

Les règles que nous venons de voir (Bayes, Probabilités Totales) fonctionnent aussi si nous avons déjà une information de base E.

Théorème : Règle de Bayes avec conditionnement additionnel

À condition que $P(A \cap E) > 0$ et $P(B \cap E) > 0$, nous avons :

$$P(A|B, E) = \frac{P(B|A, E)P(A|E)}{P(B|E)}$$

La preuve consiste à appliquer la définition de la probabilité conditionnelle à un univers déjà restreint par E.

Preuve

La preuve est identique à celle de la règle de Bayes standard, mais en appliquant la définition de la probabilité conditionnelle à un univers restreint E.

$$P(A|B, E) = P(A|(B \cap E)) = \frac{P(A \cap (B \cap E))}{P(B \cap E)}$$

$$P(B|A, E) = P(B|(A \cap E)) = \frac{P(B \cap (A \cap E))}{P(A \cap E)}$$

De la première équation : $P(A \cap B \cap E) = P(A|B, E)P(B \cap E)$. De la seconde : $P(A \cap B \cap E) = P(B|A, E)P(A \cap E)$. En égalisant : $P(A|B, E)P(B \cap E) = P(B|A, E)P(A \cap E)$. D'où : $P(A|B, E) = \frac{P(B|A,E)P(A \cap E)}{P(B \cap E)}$. En utilisant $P(X \cap Y) = P(X|Y)P(Y)$, on a $P(A \cap E) = P(A|E)P(E)$ et $P(B \cap E) = P(B|E)P(E)$.

$$P(A|B, E) = \frac{P(B|A, E)P(A|E)P(E)}{P(B|E)P(E)} = \frac{P(B|A, E)P(A|E)}{P(B|E)}$$

Cette formule peut sembler intimidante, mais elle signifie simplement que nous appliquons la même logique dans un "sous-monde" :

Intuition

Cette formule est simplement la règle de Bayes standard, mais appliquée à l'intérieur d'un univers que l'on a déjà "rétréci".
Imaginez que vous recevez une information **E** qui élimine une grande partie des possibilités. C'est votre nouveau point de départ, votre monde est plus petit. Toutes les probabilités que vous calculez désormais sont relatives à ce monde restreint.
Dans ce nouveau monde, vous recevez une autre information, l'évidence **B**. La règle de Bayes conditionnelle vous permet alors de mettre à jour votre croyance sur un événement **A**, en utilisant exactement la même logique que la règle de Bayes classique, mais en vous assurant que chaque calcul reste confiné à l'intérieur des frontières de l'univers défini par **E**.

2.7 Formule des Probabilités Totales avec Conditionnement Additionnel

De même, la loi des probabilités totales s'adapte à ce nouvel univers restreint.

Théorème : Formule des probabilités totales avec conditionnement additionnel

Soit $A_1, \ldots, A_n$ une partition de S. À condition que $P(A_i \cap E) > 0$ pour tout i, nous avons :

$$P(B|E) = \sum_{i=1}^{n} P(B|A_i, E)P(A_i|E)$$

La démonstration est une application directe de la formule standard, mais à l'intérieur de l'univers E.

Preuve

La preuve suit celle de la formule des probabilités totales standard, mais tout est conditionné par E. Soit $P_E(\cdot)$ une mesure de probabilité définie par $P_E(X) = P(X|E)$. Les A_i forment une partition de S, donc les $(A_i \cap E)$ forment une partition de E. On applique la formule standard

à $B \cap E$:

$$P(B|E) = \sum_{i=1}^{n} P(B \cap A_i|E)$$

Par la définition de la probabilité conditionnelle :

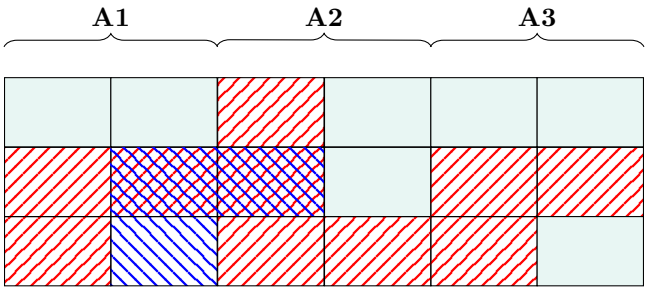
$$P(B \cap A_i|E) = \frac{P(B \cap A_i \cap E)}{P(E)}$$

Et $P(B|A_i, E)P(A_i|E) = \frac{P(B \cap A_i \cap E)}{P(A_i \cap E)} \times \frac{P(A_i \cap E)}{P(E)} = \frac{P(B \cap A_i \cap E)}{P(E)}$ Les deux termes sont égaux, donc :

$$P(B|E) = \sum_{i=1}^{n} P(B|A_i, E)P(A_i|E)$$

L'exemple visuel de la carte au trésor illustre parfaitement cette double-conditionnalité :

Intuition



Imaginez que le graphique ci-dessus représente la carte d'un trésor. La carte est partitionnée en trois grandes régions : **A1**, **A2**, et **A3**. Sur cette carte, on a identifié deux types de terrains : une **zone marécageuse** (événement E, hachures rouges) qui s'étend sur **10 parcelles**, et une **zone près d'un vieux chêne** (événement B, hachures bleues) qui couvre **3 parcelles**.
On vous donne un premier indice : "Le trésor est dans la zone marécageuse (E)". Votre univers de recherche se réduit instantanément à ces 10 parcelles rouges. Puis, on vous donne un second indice : "Le trésor est aussi près d'un chêne (B)". Votre recherche se concentre alors sur les parcelles qui sont à la fois marécageuses et proches d'un chêne (les cases violettes, $B \cap E$).
La question est : "Sachant que le trésor est dans une parcelle violette, quelle est la probabilité qu'il se trouve dans la région A2 ?". On cherche donc $P(A_2|B, E)$. La règle de Bayes nous permet de le calculer.
Calcul des termes nécessaires : D'abord, nous devons évaluer les probabilités à l'intérieur du "monde marécageux" (sachant E).
La **vraisemblance** est $P(B|A_2, E)$. En se limitant aux 4 parcelles marécageuses de la région A2, une seule est aussi près d'un chêne. Donc, $P(B|A_2, E) = 1/4$.
La **probabilité a priori** est $P(A_2|E)$. Sur les 10 parcelles marécageuses, 4 sont dans la région A2. Donc, $P(A_2|E) = 4/10$.
L'**évidence**, $P(B|E)$, est la probabilité de trouver un chêne dans l'ensemble de la zone marécageuse. On peut la calculer avec la formule des probabilités totales :

$$P(B|E) = P(B|A_1, E)P(A_1|E) + P(B|A_2, E)P(A_2|E) + P(B|A_3, E)P(A_3|E)$$

$$P(B|E) = (\frac{1}{3} \times \frac{3}{10}) + (\frac{1}{4} \times \frac{4}{10}) + (0 \times \frac{3}{10}) = \frac{1}{10} + \frac{1}{10} = \frac{2}{10}$$

Application de la règle de Bayes : Maintenant, nous assemblons le tout.

$$P(A_2|B, E) = \frac{P(B|A_2, E)P(A_2|E)}{P(B|E)} = \frac{(1/4) \times (4/10)}{2/10} = \frac{1/10}{2/10} = \frac{1}{2}$$

L'intuition confirme le calcul : sachant que le trésor est sur une parcelle violette, et qu'il n'y en a que deux (une en A1, une en A2), il y a bien une chance sur deux qu'il se trouve dans la région A2.

2.8 Indépendance de Deux Événements

Le concept d'indépendance est un cas spécial de probabilité conditionnelle où l'information B n'a aucun effet sur la probabilité de A.

Définition : Indépendance de deux événements

Les événements A et B sont indépendants si :

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

Si $P(A) > 0$ et $P(B) > 0$, cela est équivalent à :

$$P(A|B) = P(A)$$

En d'autres termes :

Intuition

L'indépendance est l'absence d'information. Si deux événements sont indépendants, apprendre que l'un s'est produit ne change absolument rien à la probabilité de l'autre. Savoir qu'il pleut à Tokyo (B) ne modifie pas la probabilité que vous obteniez pile en lançant une pièce (A).

2.9 Indépendance Conditionnelle

Attention : l'indépendance n'est pas la même chose que l'exclusion mutuelle. Il faut aussi se méfier de l'indépendance qui n'est qu'apparente, ou qui dépend d'une autre condition.

Définition : Indépendance Conditionnelle

Les événements A et B sont dits conditionnellement indépendants étant donné E si :

$$P(A \cap B|E) = P(A|E)P(B|E)$$

C'est un concept subtil mais crucial :

Intuition

L'indépendance peut apparaître ou disparaître quand on observe un autre événement. Par exemple, vos notes en maths (A) et en physique (B) ne sont probablement pas indépendantes. Mais si l'on sait que vous avez beaucoup travaillé (E), alors vos notes en maths et en physique pourraient devenir indépendantes. L'information "vous avez beaucoup travaillé" explique la corrélation ; une fois qu'on la connaît, connaître votre note en maths n'apporte plus d'information sur votre note en physique.

2.10 Le Problème de Monty Hall

Pour tester notre compréhension de tous ces concepts, le problème de Monty Hall est un exercice incontournable. Il met en lumière à quel point notre intuition sur la mise à jour des probabilités peut être faussée.

Remarque : Le problème de Monty Hall

Imaginez que vous êtes à un jeu télévisé. Face à vous se trouvent trois portes fermées. Derrière l'une d'elles se trouve une voiture, et derrière les deux autres, des chèvres.

1. Vous choisissez une porte (disons, la porte n°1).
2. L'animateur, qui sait où se trouve la voiture, ouvre une autre porte (par exemple, la n°3) derrière laquelle se trouve une chèvre.
3. Il vous demande alors : "Voulez-vous conserver votre choix initial (porte n°1) ou changer pour l'autre porte restante (la n°2) ?"

Question : Avez-vous intérêt à changer de porte ? Votre probabilité de gagner la voiture est-elle plus grande si vous changez, si vous ne changez pas, ou est-elle la même dans les deux cas ?

La réponse est contre-intuitive pour la plupart des gens, mais mathématiquement claire.

Solution du problème de Monty Hall

La réponse est sans équivoque : il faut **toujours changer de porte**. Cette stratégie fait passer la probabilité de gagner de 1/3 à 2/3. L'intuition et la preuve ci-dessous détaillent ce résultat surprenant.

Pourquoi ? L'erreur est de penser que l'animateur agit au hasard.

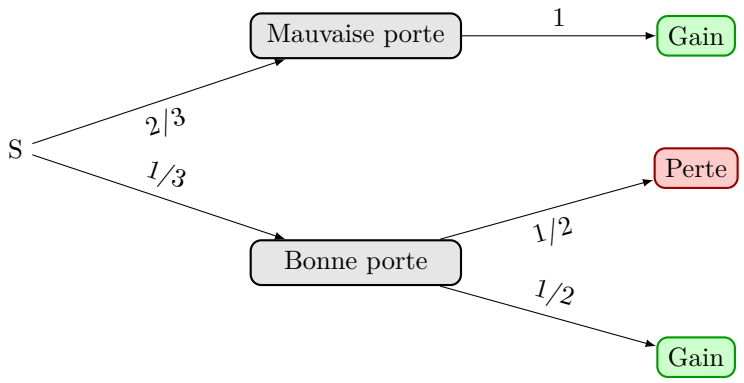
Intuition : Le secret : l'information de l'animateur

L'erreur commune est de supposer qu'il reste deux portes avec une chance égale de 1/2. Cela ignore une information capitale : le choix de l'animateur n'est **pas aléatoire**. Il sait où se trouve la voiture et ouvrira toujours une porte perdante.
Le raisonnement correct se déroule en deux temps. D'abord, votre choix initial a **1/3** de chance d'être correct. Cela implique qu'il y a **2/3** de chance que la voiture soit derrière l'une des *deux autres portes*. Ensuite, lorsque l'animateur ouvre l'une de ces deux portes, il ne fait que vous montrer où la voiture n'est *pas* dans cet ensemble. La probabilité de 2/3 se **concentre** alors entièrement sur la seule porte qu'il a laissée fermée. Changer de porte revient à miser sur cette probabilité de 2/3.

La preuve la plus claire est de suivre les stratégies :

Preuve : Preuve par l'arbre de décision

L'analyse de la meilleure stratégie peut être visualisée à l'aide de l'arbre de décision ci-dessous. Il décompose le problème en deux scénarios initiaux : avoir choisi la bonne porte (probabilité 1/3) ou une mauvaise porte (probabilité 2/3).



Analyse de l'arbre :
Branche du bas (cas le plus probable) : Avec une probabilité de **2/3**, votre choix initial se porte sur une "Mauvaise porte". L'animateur est alors obligé de révéler l'autre porte perdante. La seule porte restante est donc la bonne. L'arbre montre que cela mène à un "Gain" avec une probabilité de **1**. Ce chemin correspond au résultat de la stratégie **"Changer"**.
Branche du haut (cas le moins probable) : Avec une probabilité de **1/3**, vous avez choisi la "Bonne porte" du premier coup. L'arbre se divise alors en deux issues équiprobables (1/2 chacune). L'issue "Gain" correspond à la stratégie **"Garder"** votre choix initial, tandis que l'issue "Perte" correspond à la stratégie **"Changer"** pour la porte perdante restante.
Conclusion : Pour évaluer la meilleure stratégie, il suffit de sommer les probabilités de gain. La **probabilité de gain en changeant** est de **2/3**, car vous gagnez uniquement si votre choix initial était mauvais (branche du bas). La **probabilité de gain en gardant** est de **1/3**, car vous gagnez uniquement si votre choix initial était bon (branche "Gain" du haut). La stratégie optimale est donc bien de toujours changer de porte.

3 Variables Aléatoires Discrètes

3.1 Variable Aléatoire

Jusqu'à présent, nous avons parlé d'événements (comme "obtenir Pile" ou "tirer un Roi"). Pour analyser ces phénomènes avec des outils mathématiques plus puissants, nous devons traduire ces résultats concrets en nombres. C'est le rôle de la variable aléatoire.

Définition : Variable Aléatoire

Étant donné une expérience avec un univers S, une variable aléatoire est une fonction de l'univers S vers les nombres réels $\mathbb{R}$.

Cette définition formelle masque une idée très simple :

Intuition

Une variable aléatoire est une manière de traduire les résultats d'une expérience en nombres. Au lieu de travailler avec des concepts comme "Pile" ou "Face", on leur assigne des valeurs numériques (par exemple, 1 pour Pile, 0 pour Face). Cela nous permet d'utiliser toute la puissance des outils mathématiques (fonctions, calculs, etc.) pour analyser le hasard. C'est un pont entre le monde concret des événements et le monde abstrait des nombres.

Prenons un exemple classique :

Exemple

On lance deux dés. L'univers S est l'ensemble des 36 paires de résultats, comme $(1,1),(1,2),\ldots,(6,6)$. On peut définir une variable aléatoire X comme étant la **somme des deux dés**. Pour le résultat $(2,5)$, la valeur de la variable aléatoire est $X(2,5) = 2+5 = 7$.

3.2 Variable Aléatoire Discrète

Les variables aléatoires peuvent être de différents types. Nous commençons par le type le plus simple à "compter".

Définition : Variable Aléatoire Discrète

Une variable aléatoire X est dite discrète s'il existe une liste finie ou infinie dénombrable de valeurs $a_1, a_2, \ldots$ telle que $P(X = a_j \text{ pour un certain } j) = 1$.

L'analogie la plus simple pour comprendre le terme "discret" est celle d'un escalier.

Intuition

Une variable aléatoire est "discrète" si on peut lister (compter) toutes les valeurs qu'elle peut prendre, même si cette liste est infinie. Pensez aux "sauts" d'une valeur à l'autre, sans possibilité de prendre une valeur intermédiaire. C'est comme monter un escalier : on peut être sur la marche 1, 2 ou 3, mais jamais sur la marche 2.5. Le nombre de têtes en 10 lancers, le résultat d'un dé, le nombre d'emails que vous recevez en une heure sont des exemples. À l'opposé, une variable continue pourrait être la taille exacte d'une personne, qui peut prendre n'importe quelle valeur dans un intervalle.

3.3 Fonction de Masse (PMF)

Maintenant que nous avons une variable aléatoire qui produit des nombres discrets, nous avons besoin d'une fonction pour décrire la probabilité de chacun de ces nombres.

Définition : Probability Mass Function (PMF)

La fonction de masse (PMF) d'une variable aléatoire discrète X est la fonction P_X donnée par $P_X(x) = P(X = x)$.

C'est la "carte d'identité" probabiliste de la variable :

Intuition
La PMF est la "carte d'identité" probabiliste d'une variable aléatoire discrète. Pour chaque valeur que la variable peut prendre, la PMF nous donne la probabilité exacte associée à cette valeur. C'est comme si chaque résultat possible avait une "étiquette de prix" qui indique sa chance de se produire. La somme de toutes ces probabilités doit bien sûr valoir 1.

Un exemple très simple est le lancer de dé :

Exemple
Soit X le résultat d'un lancer de dé équilibré. La variable X peut prendre les valeurs $\{1, 2, 3, 4, 5, 6\}$. La PMF de X est la fonction qui assigne $1/6$ à chaque valeur : $P(X = 1) = 1/6$, $P(X = 2) = 1/6$, ..., $P(X = 6) = 1/6$. Pour toute autre valeur x (par exemple $x = 2.5$ ou $x = 7$), $P(X = x) = 0$.

3.4 Loi de Bernoulli

Commençons par la loi de probabilité discrète la plus simple.

Définition : Distribution de Bernoulli
Une variable aléatoire X suit la distribution de Bernoulli avec paramètre p si $P(X = 1) = p$ et $P(X = 0) = 1 - p$, où $0 < p < 1$. On note cela $X \sim \text{Bern}(p)$.

C'est la brique fondamentale de nombreuses autres distributions.

Intuition
La distribution de Bernoulli est le modèle le plus simple pour une expérience aléatoire avec seulement deux issues : "succès" (codé par 1) et "échec" (codé par 0). C'est la brique de base de nombreuses autres distributions. Pensez à un unique lancer de pièce (Pile/Face), un unique tir au but (Marqué/Manqué), ou la réponse à une question par oui/non. Le paramètre p est simplement la probabilité du "succès".

3.5 Loi Binomiale

Que se passe-t-il si nous répétons une expérience de Bernoulli n fois et que nous comptons le nombre total de succès ?

Théorème : PMF Binomiale
Si $X \sim \text{Bin}(n, p)$, alors la PMF de X est :

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

pour $k = 0, 1, \ldots, n$.

La preuve de cette formule est un argument combinatoire direct.

Preuve
Nous voulons trouver la probabilité d'obtenir exactement k succès au cours de n essais indépendants.

1. **Probabilité d'une séquence spécifique :** Considérons d'abord une séquence spécifique contenant k succès (S) et $n - k$ échecs (E), par exemple $S, S, \ldots, S, E, E, \ldots, E$. Puisque les essais sont indépendants, la probabilité de cette séquence est le produit des probabilités individuelles :

$$\underbrace{p \times p \times \cdots \times p}_{k \text{ fois}} \times \underbrace{(1-p) \times \cdots \times (1-p)}_{n-k \text{ fois}} = p^k (1-p)^{n-k}$$

2. **Nombre de séquences possibles :** La séquence ci-dessus n'est qu'une des nombreuses façons d'obtenir k succès. Le nombre total de façons d'arranger k succès parmi n positions (essais) est donné par le coefficient binomial $\binom{n}{k}$.

3. **Probabilité totale :** Chacune de ces $\binom{n}{k}$ séquences a la même probabilité $p^k(1-p)^{n-k}$. Puisque toutes ces séquences sont des événements disjoints, la probabilité totale d'obtenir k succès (dans n'importe quel ordre) est la somme de leurs probabilités :

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

Chaque partie de cette formule a une signification logique claire.

Intuition

La distribution binomiale répond à la question : "Si je répète n fois la même expérience de Bernoulli (qui a une probabilité de succès p), quelle est la probabilité d'obtenir exactement k succès ?" La formule est construite logiquement en multipliant trois composantes. D'abord, $\mathbf{p^k}$ représente la probabilité d'obtenir k succès. Ensuite, $\mathbf{(1-p)^{n-k}}$ est la probabilité que les $n-k$ échecs restants se produisent. Finalement, comme les k succès peuvent apparaître n'importe où parmi les n essais, on multiplie par $\binom{\mathbf{n}}{\mathbf{k}}$, qui compte le nombre de manières distinctes de placer ces succès.

Appliquons cela à un exemple classique :

Exemple

On lance une pièce équilibrée 10 fois ($n = 10$, $p = 0.5$). Quelle est la probabilité d'obtenir exactement 6 Piles ($k = 6$) ?

$$P(X = 6) = \binom{10}{6} (0.5)^6 (1-0.5)^{10-6} = \frac{10!}{6!4!} (0.5)^{10} = 210 \times (0.5)^{10} \approx 0.205$$

Il y a environ 20.5% de chance d'obtenir exactement 6 Piles.

3.6 Loi Hypergéométrique

La loi binomiale suppose que les essais sont indépendants, ce qui est vrai si l'on tire *avec remise*. Que se passe-t-il si l'on tire *sans remise* ?

Théorème : PMF Hypergéométrique

Si $X \sim \text{HG}(w, b, m)$, alors la PMF de X est :

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{m-k}}{\binom{w+b}{m}}$$

La preuve de cette formule est un argument de dénombrement pur, basé sur la définition naïve de la probabilité.

Preuve

Nous utilisons la définition naïve $P(A) = |A|/|S|$. Nous tirons m boules d'une urne contenant w blanches et b noires, soit $w + b$ boules au total.

1. **Taille de l'univers ($|S|$) :** Le nombre total de façons de choisir m boules parmi $w + b$ est $\binom{w+b}{m}$.
2. **Taille de l'événement favorable ($|A|$) :** Nous voulons l'événement A = "obtenir exactement k boules blanches ET $m - k$ boules noires".
   - Le nombre de façons de choisir k blanches parmi w est $\binom{w}{k}$.
   - Le nombre de façons de choisir $m - k$ noires parmi b est $\binom{b}{m-k}$.

   Par le principe de la multiplication (dénombrement), le nombre total de façons de réaliser A est $|A| = \binom{w}{k}\binom{b}{m-k}$.
3. **Probabilité :** En divisant le nombre d'issues favorables par le nombre total d'issues, on obtient :

$$P(X = k) = \frac{|A|}{|S|} = \frac{\binom{w}{k}\binom{b}{m-k}}{\binom{w+b}{m}}$$

Chaque terme de cette fraction a un sens très concret :

Intuition

La distribution hypergéométrique est la "cousine" de la binomiale pour les tirages **sans remise**. Imaginez une urne avec des boules de deux couleurs (par exemple, w blanches et b noires). Vous tirez m boules d'un coup. Quelle est la probabilité que vous ayez exactement k boules blanches ? La formule est un simple ratio issu du dénombrement. Le **dénominateur**, $\binom{w+b}{m}$, compte le nombre total de façons de tirer m boules parmi toutes celles disponibles. Le **numérateur** compte les issues favorables : c'est le produit du nombre de façons de choisir k blanches parmi les w ($\binom{w}{k}$) ET de choisir les $m-k$ boules restantes parmi les noires ($\binom{b}{m-k}$). La différence clé avec la loi binomiale est que les tirages ne sont pas indépendants.

Un exemple typique est la formation de comités à partir d'un groupe.

Exemple

Un comité de 5 personnes est choisi au hasard parmi un groupe de 8 hommes et 10 femmes. Quelle est la probabilité que le comité soit composé de 2 hommes et 3 femmes ? Ici, on tire 5 personnes ($m = 5$) d'une population de 18 personnes. On s'intéresse au nombre d'hommes ($k = 2$) parmi les 8 disponibles ($w = 8$). Le reste du comité sera composé de femmes ($b = 10$).

$$P(X = 2) = \frac{\binom{8}{2}\binom{10}{3}}{\binom{18}{5}} = \frac{28 \times 120}{8568} \approx 0.392$$

Il y a environ 39.2% de chance que le comité ait exactement cette composition.

3.7 Loi Géométrique

Revenons aux essais de Bernoulli (indépendants). Au lieu de fixer le nombre d'essais n, demandons-nous : combien d'essais faut-il avant d'obtenir notre premier succès ?

Théorème : PMF de la loi géométrique

Une variable aléatoire X suit la loi géométrique de paramètre p, notée $X \sim \text{Geom}(p)$, si elle modélise le nombre d'échecs avant le premier succès dans une série d'épreuves de Bernoulli indépendantes. Sa fonction de masse (PMF) est :

$$P(X = k) = (1-p)^k p \quad \text{pour } k = 0, 1, 2, \ldots$$

où $q = 1 - p$ est la probabilité d'échec.

La preuve de cette formule est une application directe de l'indépendance des essais.

Preuve

Soit S_i l'événement "succès au i-ème essai" et E_i l'événement "échec au i-ème essai". L'événement $\{X = k\}$ signifie que nous avons observé exactement k échecs, suivis d'un succès au $(k+1)$-ème essai. C'est la séquence d'événements : $E_1 \cap E_2 \cap \cdots \cap E_k \cap S_{k+1}$.
Puisque tous les essais sont indépendants, la probabilité de cette intersection est le produit des probabilités individuelles :

$$\begin{aligned} P(X = k) &= P(E_1) \times P(E_2) \times \cdots \times P(E_k) \times P(S_{k+1}) \\ &= \underbrace{(1-p) \times (1-p) \times \cdots \times (1-p)}_{k \text{ fois}} \times p \\ &= (1-p)^k p \end{aligned}$$

La formule est donc très littérale :

Intuition

La formule $P(X = k) = q^k p$ décrit la probabilité d'une séquence très spécifique : k échecs consécutifs (chacun avec une probabilité q, donc q^k pour la série), suivis immédiatement d'un succès (avec une probabilité p). C'est la loi de "l'attente du premier succès".

Un exemple classique est l'attente d'un résultat spécifique sur un dé.

Exemple : Premier 6 au lancer de dé

On lance un dé jusqu'à obtenir un 6. La probabilité de succès est $p = 1/6$, et celle d'échec est $q = 5/6$. Quelle est la probabilité que l'on ait besoin de 3 lancers (donc 2 échecs avant le premier succès) ? Ici, $k = 2$. La probabilité est :

$$P(X = 2) = (5/6)^2 \cdot (1/6) = \frac{25}{216} \approx 0.116$$

3.8 Loi de Poisson

Introduisons maintenant une loi utilisée pour modéliser le nombre d'événements se produisant dans un intervalle de temps ou d'espace fixe.

Définition : Distribution de Poisson

Une variable aléatoire X suit la loi de Poisson de paramètre $\lambda > 0$ si sa PMF est donnée par :

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!} \quad \text{pour } k = 0, 1, 2, \dots$$

Elle modélise typiquement le nombre d'événements se produisant dans un intervalle de temps ou d'espace fixe.

Cette loi est souvent appelée la loi des événements rares.

Intuition

La loi de Poisson est la loi des événements rares. Imaginez que vous comptez le nombre d'appels arrivant à un standard téléphonique en une minute. Il y a de nombreux instants où un appel pourrait arriver, mais la probabilité à chaque instant est infime. La loi de Poisson modélise ce type de scénario, où l'on connaît seulement le taux moyen d'arrivée des événements (λ).

Mais d'où vient cette formule avec e et une factorielle ? Elle vient d'une approximation de la loi binomiale lorsque n est très grand et p très petit.

Théorème : La loi de Poisson comme limite de la loi binomiale

Soit $X_n \sim \text{Bin}(n, p_n)$, où $\lambda = np_n$ est une constante positive fixée. Alors, pour tout $k \in \{0, 1, 2, \dots\}$, nous avons :

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

En pratique, la loi de Poisson est une excellente approximation de la loi binomiale quand n est grand et p est petit.

Intuition : Convergence Binomiale vers Poisson : L'Exemple des Naissances

Supposons que les bébés naissent dans une grande ville à un taux moyen de $\lambda = 10$ naissances par jour. Comment modéliser le nombre X de naissances un jour donné ?

1. Approche Binomiale (Découpage du Temps) : On peut diviser la journée (24h) en n très petits intervalles de temps (par exemple, $n = 24 \times 60 \times 60 = 86400$ secondes).

- Si n est très grand, la chance p qu'une naissance se produise *exactement* pendant une seconde donnée est minuscule. On peut calculer cette probabilité p comme le taux moyen divisé par le nombre d'intervalles : $p = \lambda/n = 10/86400$.
- On peut aussi supposer que la probabilité d'avoir *deux* naissances ou plus dans la même seconde est négligeable. Chaque seconde est donc comme un mini-essai de Bernoulli : soit 1 naissance (avec probabilité p), soit 0 naissance (avec probabilité $1 - p$).
- Le nombre total de naissances X sur la journée est la somme de ces n essais de Bernoulli quasi-indépendants. X suit donc approximativement une loi binomiale : $X \approx \text{Bin}(n, p = \lambda/n)$.

La probabilité d'avoir exactement k naissances serait $P(X = k) \approx \binom{n}{k} p^k (1 - p)^{n-k}$.

2. Le Passage à la Limite (Modèle Continu) : Que se passe-t-il si on rend les intervalles de temps infiniment petits ($n \to \infty$) ? C'est là que la magie opère :

- Le terme $\binom{n}{k}$ (combien de façons de choisir k secondes parmi n) se comporte comme $n^k/k!$ pour n grand.
- Le terme $p^k = (\lambda/n)^k$ devient λ^k/n^k.
- Le terme $(1-p)^{n-k} = (1-\lambda/n)^{n-k}$. Comme k est petit par rapport à n, ceci est très proche de $(1-\lambda/n)^n$, qui tend vers $e^{-\lambda}$.

En combinant ces approximations (expliquées plus en détail dans la preuve formelle), on trouve que la probabilité $P(X=k)$ tend vers $\frac{n^k}{k!}\frac{\lambda^k}{n^k}e^{-\lambda} = \frac{e^{-\lambda}\lambda^k}{k!}$.
Conclusion : La loi de Poisson apparaît naturellement comme la limite d'un processus binomial où l'on a un très grand nombre d'opportunités (n) pour qu'un événement rare (probabilité p) se produise, tout en maintenant un taux moyen constant ($\lambda = np$).

La preuve formelle montre comment les termes de la formule binomiale se transforment en ceux de la formule de Poisson lorsque $n \to \infty$.

Preuve : Dérivation de la loi de Poisson à partir de la loi Binomiale (Détaillée)

On part de la fonction de masse (PMF) d'une variable aléatoire X_n suivant une loi binomiale $\text{Bin}(n,p)$, où l'on pose $p = \lambda/n$. L'objectif est de trouver la limite de cette PMF lorsque n tend vers l'infini, tout en gardant $\lambda = np$ constant (ce qui implique que p doit tendre vers 0).
La PMF binomiale est :

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

Substituons $p = \lambda/n$:

$$P(X_n = k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

Maintenant, développons le coefficient binomial $\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}$:

$$P(X_n = k) = \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

Réorganisons les termes pour isoler ceux qui dépendent de n :

$$P(X_n = k) = \frac{\lambda^k}{k!} \times \frac{n(n-1)\cdots(n-k+1)}{n^k} \times \left(1 - \frac{\lambda}{n}\right)^n \times \left(1 - \frac{\lambda}{n}\right)^{-k}$$

Nous allons maintenant examiner la limite de chaque partie lorsque $n \to \infty$, pour k et λ fixés.

1. $\frac{\lambda^k}{k!}$: Ce terme est constant par rapport à n, donc sa limite est lui-même.
2. $\frac{n(n-1)\cdots(n-k+1)}{n^k}$: Ce terme est un produit de k facteurs divisé par n^k. On peut le réécrire comme :
$$\frac{n}{n} \times \frac{n-1}{n} \times \frac{n-2}{n} \times \cdots \times \frac{n-k+1}{n}$$
$$= 1 \times \left(1 - \frac{1}{n}\right) \times \left(1 - \frac{2}{n}\right) \times \cdots \times \left(1 - \frac{k-1}{n}\right)$$
Lorsque $n \to \infty$, chacun des termes $\frac{1}{n}, \frac{2}{n}, \ldots, \frac{k-1}{n}$ tend vers 0 (car k est fixe). Donc, chaque parenthèse tend vers $(1-0) = 1$. Puisqu'il y a un nombre *fixe* k de termes dans le produit, la limite du produit est le produit des limites :
$$\lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{n^k} = 1 \times 1 \times \cdots \times 1 = 1$$
Intuition : Pour n très grand par rapport à k, les k termes $n, n-1, \ldots, n-k+1$ sont tous "presque" égaux à n. Leur produit est donc "presque" n^k, et le ratio est "presque" 1.
3. $\left(1 - \frac{\lambda}{n}\right)^n$: C'est une limite fondamentale en analyse. On sait que pour tout réel x, $\lim_{n\to\infty}(1+x/n)^n = e^x$. Ici, nous avons $x = -\lambda$. Donc :
$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}$$

Intuition : C'est la définition même de l'exponentielle comme limite d'intérêts composés continus (ici, avec un taux négatif).

4. $\mathbf{\left(1 - \frac{\lambda}{n}\right)^{-k}}$: Lorsque $n \to \infty$, le terme λ/n tend vers 0. L'expression à l'intérieur de la parenthèse tend donc vers $(1 - 0) = 1$. Puisque k est un exposant fixe :

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-k} = 1^{-k} = 1$$

Intuition : Pour n très grand, $(1-\lambda/n)$ est très proche de 1. Élever ce nombre très proche de 1 à une puissance fixe k le laisse très proche de 1.

Finalement, en multipliant les limites de chaque partie (puisque la limite d'un produit est le produit des limites) :

$$\lim_{n\to\infty} P(X_n = k) = \left(\frac{\lambda^k}{k!}\right) \times (1) \times (e^{-\lambda}) \times (1) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Ceci est exactement la fonction de masse de probabilité d'une loi de Poisson de paramètre λ.

Un ensemble de données historiques célèbres illustre parfaitement cette loi.

Exemple : Décès par ruade de cheval : Les données de Bortkiewicz

En 1898, le statisticien Ladislaus Bortkiewicz a publié des données célèbres sur le nombre de soldats de la cavalerie prussienne tués par des ruades de cheval. Ces données sont un exemple classique d'application de la loi de Poisson pour modéliser des événements rares.

Contexte et calcul du paramètre λ : Sur une période de 20 ans, en observant 10 corps d'armée, il a collecté des données sur 200 "corps-années". Durant cette période, il y a eu un total de 122 décès. Le taux moyen de décès par corps-année est donc :

$$\lambda = \frac{\text{Nombre total de décès}}{\text{Nombre total de corps-années}} = \frac{122}{200} = 0.61$$

Le nombre de décès par corps-année, X, est donc modélisé par une loi de Poisson : $X \sim \text{Poisson}(\lambda = 0.61)$.

Comparaison des données observées et des prédictions du modèle : On peut calculer la probabilité d'observer k décès en une année-corps en utilisant la PMF de Poisson : $P(X = k) = \frac{e^{-0.61}(0.61)^k}{k!}$. En multipliant cette probabilité par le nombre total d'observations (200), on obtient le nombre de cas attendus (nombre de corps d'armes dans lequels il y a k deces).

Nombre de décès (k)	Observé	Probabilité de Poisson	Attendu
0	109	$P(X = 0) \approx 0.543$	108.7
1	65	$P(X = 1) \approx 0.331$	66.3
2	22	$P(X = 2) \approx 0.101$	20.2
3	3	$P(X = 3) \approx 0.021$	4.1
4	1	$P(X = 4) \approx 0.003$	0.6
5+	0	$P(X \geq 5) \approx 0.000$	0.0

L'adéquation remarquable entre les fréquences observées et les valeurs attendues par le modèle de Poisson a contribué à populariser cette distribution pour l'analyse d'événements rares.

3.9 Fonction de Répartition (CDF)

Nous avons la PMF, qui donne $P(X = x)$. Une autre fonction tout aussi importante est la fonction de répartition (CDF), qui "accumule" ces probabilités.

Définition : Cumulative Distribution Function (CDF)

La fonction de répartition (CDF) d'une variable aléatoire X est la fonction F_X donnée par $F_X(x) = P(X \leq x)$.

Cette fonction répond à une question différente de celle de la PMF.

Intuition
Alors que la PMF répond à la question "Quelle est la probabilité d'obtenir *exactement* x ?", la CDF répond à la question "Quelle est la probabilité d'obtenir *au plus* x ?". C'est une fonction cumulative : pour une valeur x donnée, elle additionne les probabilités de tous les résultats inférieurs ou égaux à x. La CDF a toujours une forme d'escalier pour les variables discrètes. Elle commence à 0 (très loin à gauche) et monte par "sauts" à chaque valeur possible de la variable, pour finalement atteindre 1 (très loin à droite). La hauteur de chaque saut correspond à la valeur de la PMF à ce point.

Traçons cette fonction "en escalier" pour notre exemple du dé.

Exemple

Reprenons le lancer d'un dé équilibré (X). Calculons quelques valeurs de la CDF, notée $F(x)$.
$F(0.5) = P(X \leq 0.5) = 0$
$F(1) = P(X \leq 1) = P(X = 1) = 1/6$
$F(1.5) = P(X \leq 1.5) = P(X = 1) = 1/6$
$F(2) = P(X \leq 2) = P(X = 1) + P(X = 2) = 2/6$
$F(5.9) = P(X \leq 5.9) = P(X = 1) + \cdots + P(X = 5) = 5/6$
$F(6) = P(X \leq 6) = 1$
$F(100) = P(X \leq 100) = 1$

3.10 Variable Aléatoire Indicatrice

Enfin, nous introduisons un outil simple mais qui s'avérera extraordinairement puissant pour les preuves, notamment celles concernant l'espérance.

Définition : Variable Aléatoire Indicatrice
La variable aléatoire indicatrice d'un événement A est la variable aléatoire qui vaut 1 si A se produit et 0 sinon. Nous la noterons I_A. Notez que $I_A \sim \text{Bern}(p)$ avec $p = P(A)$.

C'est un simple interrupteur "on/off".

Intuition
Une variable indicatrice est un interrupteur. Elle est sur "ON" (valeur 1) si un événement qui nous intéresse se produit, et sur "OFF" (valeur 0) sinon. C'est un outil extrêmement puissant car il transforme les questions sur les probabilités des événements en questions sur les espérances des variables aléatoires, ce qui simplifie souvent les calculs.

4 Variables Aléatoires Continues

4.1 Fonction de Densité de Probabilité (PDF)

Nous passons maintenant aux variables aléatoires qui peuvent prendre n'importe quelle valeur dans un intervalle, comme la taille d'une personne ou le temps d'attente exact. Pour ces variables, la notion de PMF n'a plus de sens, car la probabilité d'obtenir une valeur *exacte* est nulle. Nous introduisons donc le concept de densité.

Définition : Fonction de Densité de Probabilité (PDF)

Soit X une variable aléatoire continue. Une fonction f est une **fonction de densité de probabilité** (Probability Density Function, ou PDF) de X si, pour tout x :

1. $f(x) \geq 0$, pour tout $-\infty < x < \infty$
2. $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$ (l'aire totale sous la courbe vaut 1)

Il est crucial de comprendre que $f(x)$ n'est *pas* une probabilité.

Intuition

Dans le cas discret, la PMF donnait une "masse" de probabilité à chaque point. Dans le cas continu, la probabilité en un point exact est nulle ($P(X = x) = 0$). La PDF, $f(x)$, n'est **pas** une probabilité.
Il faut voir $f(x)$ comme une **densité** : elle décrit la "concentration" de probabilité autour de x. Pour obtenir une probabilité (une "masse"), il faut intégrer cette densité sur un intervalle. La probabilité que X tombe dans un intervalle $[a, b]$ est l'aire sous la courbe de la PDF entre a et b :

$$P(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x$$

Cette distinction est fondamentale.

Remarque : PDF vs Probabilité

Une erreur fréquente est de confondre la valeur $f(x)$ avec $P(X = x)$. Pour une variable continue, $P(X = x)$ est **toujours zéro**. La PDF $f(x)$ peut être supérieure à 1 (contrairement à une probabilité), tant que l'aire totale sous la courbe reste égale à 1. Pensez-y comme à une densité de population : elle peut être très élevée en un point, mais la "population" (probabilité) exacte en ce point infinitésimal est nulle.

Vérifions un exemple simple.

Exemple : Une PDF simple

Soit X une v.a. avec la PDF $f(x) = 2x$ pour $x \in [0, 1]$, et $f(x) = 0$ sinon.

1. Est-ce une PDF valide ?
   (1) $f(x) \geq 0$ pour tout x dans $[0, 1]$.
   (2) $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_0^1 2x\,\mathrm{d}x = [x^2]_0^1 = 1 - 0 = 1$.
   Oui, c'est une PDF valide.
2. Quelle est la probabilité $P(X \leq 0.5)$?

$$P(X \leq 0.5) = \int_0^{0.5} 2x\,\mathrm{d}x = [x^2]_0^{0.5} = (0.5)^2 - 0 = 0.25$$

4.2 Fonction de Répartition (CDF)

Comme dans le cas discret, nous pouvons définir une fonction qui accumule la probabilité. Pour le cas continu, cette accumulation se fait via une intégrale.

Définition : Fonction de Répartition Continue (CDF)

Soit X une variable aléatoire continue. La **fonction de répartition** (Cumulative Distribution Function, ou CDF) de X est la fonction F définie par :

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$$

Pour être une CDF valide, la fonction F doit respecter les propriétés suivantes :

1. $\lim_{x\to\infty} F(x) = 1$
2. $\lim_{x\to-\infty} F(x) = 0$
3. F est continue et non décroissante.

La CDF est l'intégrale de la PDF, et inversement, la PDF est la dérivée de la CDF.

Intuition

La CDF est "l'accumulateur" de probabilité. Elle part de 0 (à $-\infty$) et "accumule" l'aire sous la PDF à mesure qu'on avance sur l'axe des x, pour finalement atteindre 1 (à $+\infty$).
Le lien fondamental est que la PDF est la dérivée de la CDF (par le théorème fondamental de l'analyse) :

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

Cela signifie que la valeur de la PDF $f(x)$ représente le **taux d'accumulation** de la probabilité au point x.

La CDF est souvent le moyen le plus simple de calculer des probabilités sur des intervalles.

Remarque : Calcul de Probabilités via la CDF

La CDF est très pratique pour calculer des probabilités sur des intervalles :

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

Pour les variables continues, les inégalités strictes ou larges ne changent rien ($P(X = a) = 0$).

Calculons la CDF de notre exemple précédent.

Exemple : CDF de l'exemple précédent

Pour $f(x) = 2x$ sur $[0, 1]$, la CDF $F(x)$ est :

- Si $x < 0$: $F(x) = \int_{-\infty}^{x} 0\,\mathrm{d}t = 0$.
- Si $0 \leq x \leq 1$: $F(x) = \int_{-\infty}^{0} f(t)\mathrm{d}t + \int_{0}^{x} 2t\,\mathrm{d}t = 0 + [t^2]_0^x = x^2$.
- Si $x > 1$: $F(x) = \int_{-\infty}^{1} f(t)\mathrm{d}t + \int_{1}^{x} 0\,\mathrm{d}t = \int_{0}^{1} 2t\,\mathrm{d}t = 1$.

Donc, $F(x) = \begin{cases} 0 & \text{si } x < 0 \\ x^2 & \text{si } 0 \leq x \leq 1 \\ 1 & \text{si } x > 1 \end{cases}$

4.3 Espérance et Variance (Cas Continu)

Les concepts d'espérance et de variance s'étendent naturellement au cas continu, en remplaçant les sommes par des intégrales.

Définition : Espérance et Variance (Cas Continu)

Pour une variable aléatoire X de fonction de densité f :
L'**espérance** de X est le centre de gravité de la densité :

$$E[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$$

La **variance** de X est l'espérance du carré de l'écart à la moyenne :

$$\mathrm{Var}(X) = E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x - E[X])^2 f(x)\,\mathrm{d}x$$

Comme dans le cas discret, une formule alternative existe pour la variance.

Théorème : Formule de calcul de la Variance
Une formule plus simple pour le calcul de la variance est :

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2$$

où $E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x$. (Ceci est une application de LOTUS).

La preuve est identique à celle du cas discret, en utilisant la linéarité de l'espérance.

Preuve
Soit $\mu = E(X)$. On part de la définition de la variance :

$$\begin{aligned}
\mathrm{Var}(X) &= E[(X-\mu)^2] \\
&= E[X^2 - 2X\mu + \mu^2] \quad \text{(On développe le carré)} \\
&= E(X^2) - E(2\mu X) + E(\mu^2) \quad \text{(Par linéarité de l'espérance, qui s'applique aussi au cas continu)} \\
&= E(X^2) - 2\mu E(X) + \mu^2 \quad \text{(Car } 2\mu \text{ et } \mu^2 \text{ sont des constantes)} \\
&= E(X^2) - 2\mu(\mu) + \mu^2 \quad \text{(Car } E(X) = \mu\text{)} \\
&= E(X^2) - 2\mu^2 + \mu^2 \\
&= E(X^2) - \mu^2 = E(X^2) - [E(X)]^2
\end{aligned}$$

Le calcul de $E[X^2]$ (et plus généralement de $E[g(X)]$) repose sur le théorème de transfert, adapté au cas continu.

Théorème : Théorème de Transfert (LOTUS)
Si X est une v.a. continue de densité $f(x)$, et g une fonction, alors :

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\,\mathrm{d}x$$

La preuve formelle est plus avancée, mais l'idée est analogue au cas discret : on pondère chaque valeur $g(x)$ par la densité de probabilité $f(x)$ au voisinage de x.

Preuve : Idée de la preuve
La preuve formelle repose sur la théorie de la mesure ou sur un argument de changement de variable pour l'intégrale, en passant par la fonction de répartition de $Y = g(X)$. Intuitivement, pour un petit intervalle dx autour de x, la "masse" de probabilité est $f(x)dx$. Cette masse correspond à une valeur $g(x)$ pour la nouvelle variable. L'espérance est la somme (intégrale) de ces valeurs pondérées par leur masse : $\int g(x)f(x)dx$.

La propriété la plus importante de l'espérance reste valide.

Remarque : Linéarité de l'Espérance
Comme dans le cas discret, l'espérance reste linéaire pour les variables continues : $E[aX+bY] = aE[X] + bE[Y]$.

Calculons l'espérance et la variance pour notre exemple.

Exemple : Espérance et Variance de l'exemple précédent
Pour $f(x) = 2x$ sur $[0, 1]$:
$E[X] = \int_0^1 x \cdot (2x)\,\mathrm{d}x = \int_0^1 2x^2\,\mathrm{d}x = \left[\frac{2x^3}{3}\right]_0^1 = \frac{2}{3}$.

$E[X^2] = \int_0^1 x^2 \cdot (2x)\,\mathrm{d}x = \int_0^1 2x^3\,\mathrm{d}x = \left[\frac{2x^4}{4}\right]_0^1 = \frac{1}{2}$.
$\mathrm{Var}(X) = E[X^2] - (E[X])^2 = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{2} - \frac{4}{9} = \frac{9-8}{18} = \frac{1}{18}$.

4.4 Loi Uniforme

La loi continue la plus simple est celle où la densité est constante sur un intervalle.

Définition : Loi Uniforme
Une variable aléatoire X est **uniformément distribuée** sur un intervalle $[a, b]$ si sa densité est une constante sur cet intervalle. Pour que l'aire totale soit 1, cette constante doit être $\frac{1}{b-a}$.

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{pour } x \in [a, b] \\ 0 & \text{sinon} \end{cases}$$

On note cela $X \sim \mathrm{Unif}(a, b)$.

C'est le modèle du "hasard pur" sur un segment.

Intuition
C'est la distribution du "hasard pur" dans un intervalle borné. La probabilité de tomber dans un sous-intervalle ne dépend que de la **longueur** de ce sous-intervalle, pas de sa position (tant qu'il est dans $[a, b]$).

Les propriétés de cette loi sont faciles à dériver par intégration directe.

Théorème : Propriétés de la Loi Uniforme
Si $X \sim \mathrm{Unif}(a, b)$:

- **CDF :** $F(x) = \frac{x-a}{b-a}$ pour $x \in [a, b]$.
- **Espérance :** $E[X] = \frac{a+b}{2}$ (le point milieu de l'intervalle).
- **Variance :** $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$.

Preuve : Dérivation des propriétés
Soit $f(x) = \frac{1}{b-a}$ pour $x \in [a, b]$ et 0 sinon.
CDF : Pour $x \in [a, b]$,

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \int_a^x \frac{1}{b-a}\,dt = \frac{1}{b-a}[t]_a^x = \frac{x-a}{b-a}$$

(Pour $x < a$, $F(x) = 0$. Pour $x > b$, $F(x) = 1$.)
Espérance :

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

Variance : D'abord, calculons $E[X^2]$.

$$E[X^2] = \int_a^b x^2\frac{1}{b-a}\,dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{1}{b-a}\frac{b^3-a^3}{3}$$

En utilisant $b^3-a^3 = (b-a)(b^2+ab+a^2)$, on obtient $E[X^2] = \frac{b^2+ab+a^2}{3}$. Maintenant, appliquons

la formule $\text{Var}(X) = E[X^2] - (E[X])^2$:

$$\begin{aligned}\text{Var}(X) &= \frac{b^2+ab+a^2}{3} - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{b^2+ab+a^2}{3} - \frac{a^2+2ab+b^2}{4} \\ &= \frac{4(b^2+ab+a^2) - 3(a^2+2ab+b^2)}{12} \\ &= \frac{4b^2+4ab+4a^2-3a^2-6ab-3b^2}{12} \\ &= \frac{b^2-2ab+a^2}{12} = \frac{(b-a)^2}{12}\end{aligned}$$

4.5 Loi Exponentielle

Passons à une loi fondamentale pour modéliser les temps d'attente.

Définition : Loi Exponentielle

Une variable aléatoire X suit une **loi exponentielle** de paramètre $\lambda > 0$ si sa fonction de densité a la forme :

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{pour } x \geq 0 \\ 0 & \text{sinon} \end{cases}$$

On note $X \sim \text{Exp}(\lambda)$.

Cette loi est intimement liée au processus de Poisson.

Intuition : Lien entre les lois de Poisson et Exponentielle

La loi exponentielle modélise le temps d'attente *avant* le prochain événement dans un processus de Poisson.
Posons la question : « Si je commence à observer maintenant, combien de temps T vais-je devoir attendre avant de voir le prochain événement ? »

1. Dans un processus de Poisson de taux λ, le nombre d'événements $N(t)$ dans un intervalle de temps t suit une loi de Poisson de paramètre λt :

$$P(N(t) = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

2. La probabilité de ne voir **aucun** événement ($k = 0$) pendant une durée t est :

$$P(N(t) = 0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}$$

3. Mais ne voir aucun événement pendant un temps t, c'est exactement dire que le temps d'attente T du premier événement est *plus grand* que t.

$$P(T > t) = P(N(t) = 0) = e^{-\lambda t}$$

4. À partir de là, on déduit la fonction de répartition (CDF) de T :

$$F_T(t) = P(T \leq t) = 1 - P(T > t) = 1 - e^{-\lambda t} \quad (\text{pour } t \geq 0)$$

5. En dérivant la CDF pour obtenir la densité (PDF) :

$$f_T(t) = F_T'(t) = \frac{d}{dt}(1 - e^{-\lambda t}) = -(-\lambda e^{-\lambda t}) = \lambda e^{-\lambda t}$$

C'est exactement la densité de la loi exponentielle de paramètre λ.

Cette loi possède des propriétés remarquables.

Théorème : Propriétés de la Loi Exponentielle

Si $X \sim \text{Exp}(\lambda)$:

- **CDF :** $F(x) = 1 - e^{-\lambda x}$ pour $x \geq 0$.
- **Espérance :** $E[X] = \frac{1}{\lambda}$.
- **Variance :** $\text{Var}(X) = \frac{1}{\lambda^2}$.
- **Propriété de non-mémoire :** Pour $s, t \geq 0$, $P(X > s + t \mid X > s) = P(X > t)$.

Les preuves de l'espérance et de la variance nécessitent une intégration par parties. La preuve de la non-mémoire est plus directe.

Preuve : Dérivation des propriétés

Soit $f(x) = \lambda e^{-\lambda x}$ pour $x \geq 0$.
CDF : A été dérivée dans l'intuition ci-dessus.

$$F(x) = \int_0^x \lambda e^{-\lambda t}\, dt = [-e^{-\lambda t}]_0^x = -e^{-\lambda x} - (-e^0) = 1 - e^{-\lambda x}$$

Espérance : On utilise l'intégration par parties ($\int u dv = uv - \int v du$) avec $u = x$ et $dv = \lambda e^{-\lambda x} dx$. Alors $du = dx$ et $v = -e^{-\lambda x}$.

$$\begin{aligned} E[X] &= \int_0^\infty x(\lambda e^{-\lambda x})\, dx \\ &= \left[x(-e^{-\lambda x})\right]_0^\infty - \int_0^\infty (-e^{-\lambda x})\, dx \\ &= (0 - 0) + \int_0^\infty e^{-\lambda x}\, dx \quad (\text{car } \lim_{x\to\infty} -xe^{-\lambda x} = 0) \\ &= \left[-\frac{1}{\lambda} e^{-\lambda x}\right]_0^\infty = (0) - (-\frac{1}{\lambda} e^0) = \frac{1}{\lambda} \end{aligned}$$

Variance : D'abord $E[X^2]$. Intégration par parties avec $u = x^2$, $dv = \lambda e^{-\lambda x} dx$. $du = 2x dx$, $v = -e^{-\lambda x}$.

$$\begin{aligned} E[X^2] &= \int_0^\infty x^2(\lambda e^{-\lambda x})\, dx \\ &= \left[x^2(-e^{-\lambda x})\right]_0^\infty - \int_0^\infty (-e^{-\lambda x})(2x\, dx) \\ &= 0 + \int_0^\infty 2xe^{-\lambda x}\, dx \\ &= \frac{2}{\lambda} \int_0^\infty x(\lambda e^{-\lambda x})\, dx \quad (\text{On fait apparaître } E[X]) \\ &= \frac{2}{\lambda} E[X] = \frac{2}{\lambda}\left(\frac{1}{\lambda}\right) = \frac{2}{\lambda^2} \end{aligned}$$

Donc, $\text{Var}(X) = E[X^2] - (E[X])^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$.
Propriété de non-mémoire : Rappelons que $P(X > t) = e^{-\lambda t}$.

$$\begin{aligned} P(X > s + t \mid X > s) &= \frac{P(X > s + t \text{ et } X > s)}{P(X > s)} \\ &= \frac{P(X > s + t)}{P(X > s)} \quad (\text{car si } X > s + t, \text{ alors } X > s) \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = \frac{e^{-\lambda s} e^{-\lambda t}}{e^{-\lambda s}} = e^{-\lambda t} \\ &= P(X > t) \end{aligned}$$

Le paramètre λ a une interprétation concrète.

Remarque : Interprétation du paramètre λ

Le paramètre λ représente le **taux** moyen d'occurrence des événements dans le processus de Poisson sous-jacent (par exemple, nombre moyen d'appels par minute). L'espérance $1/\lambda$ est alors le **temps moyen entre les événements**.

La propriété de non-mémoire est unique à la loi exponentielle (dans le cas continu).

Intuition : La Propriété de Non-Mémoire

C'est la propriété la plus contre-intuitive et la plus importante de la loi exponentielle. Elle signifie que le processus "oublie" le passé. Si vous attendez un bus qui arrive selon un processus de Poisson (et donc le temps d'attente suit une loi exponentielle), et que vous avez déjà attendu 5 minutes ($X > 5$), la probabilité que vous deviez attendre encore au moins 2 minutes ($X > 5 + 2$) est la même que si vous veniez juste d'arriver à l'arrêt et deviez attendre au moins 2 minutes ($X > 2$). L'information "j'ai déjà attendu 5 minutes" est inutile pour prédire l'attente future.

4.6 Distributions Conjointes (Cas Continu)

Comme pour le cas discret, nous pouvons définir des lois conjointes pour plusieurs variables aléatoires continues.

Définition : Fonction de Densité Conjointe

Pour des variables aléatoires continues X et Y, la **fonction de densité conjointe** $f(x, y)$ décrit la densité de probabilité sur le plan (x, y). Elle doit respecter :

1. $f(x, y) \geq 0$, pour tous x, y.
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, \mathrm{d}x\, \mathrm{d}y = 1$.

Ici, la probabilité est associée à un volume sous la surface de densité.

Intuition : Volume = Probabilité

La probabilité que le couple (X, Y) tombe dans une région A du plan xy est le **volume** sous la surface $z = f(x, y)$ au-dessus de cette région A.

$$P((X, Y) \in A) = \iint_A f(x, y)\, \mathrm{d}A$$

On retrouve les lois marginales en intégrant (en "écrasant" le volume).

Définition : Densités Marginales

On peut retrouver les densités individuelles (marginales) en "écrasant" le volume 3D sur un seul axe. Pour obtenir la PDF de X seul, on intègre $f(x, y)$ sur toutes les valeurs possibles de y :

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, \mathrm{d}y$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, \mathrm{d}x$$

La CDF conjointe accumule ce volume.

Définition : CDF Conjointe

La **fonction de répartition conjointe** (CDF) est :

$$F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(s, t)\, \mathrm{d}s\, \mathrm{d}t$$

Elle représente le volume "au sud-ouest" du point (x, y).

4.7 Espérance, Indépendance et Covariance (Cas Conjoint)

Les concepts clés s'étendent naturellement au cas conjoint continu.

Théorème : LOTUS pour les v.a. conjointes

Si X et Y ont une densité conjointe $f(x, y)$, et $g(x, y)$ est une fonction :

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) \, \mathrm{d}x \, \mathrm{d}y$$

La preuve est analogue à celle de LOTUS 1D, mais en dimension supérieure.

Preuve : Idée de la preuve

Comme pour LOTUS 1D, la preuve rigoureuse utilise des arguments de théorie de la mesure. L'intuition est que pour un petit rectangle $dxdy$ autour de (x, y), la "masse" de probabilité est $f(x, y)dxdy$. Cette masse correspond à la valeur $g(x, y)$. L'espérance est la somme (double intégrale) de ces valeurs $g(x, y)$ pondérées par leur masse $f(x, y)dxdy$.

La condition d'indépendance s'exprime via la factorisation de la densité.

Définition : Indépendance et Densité

Les variables aléatoires continues X et Y sont **indépendantes** si et seulement si leur densité conjointe est le produit de leurs densités marginales :

$$f(x, y) = f_X(x) f_Y(y), \quad \text{pour tous } x, y$$

Cela signifie que le profil selon x ne dépend pas de y.

Intuition

Intuitivement, l'indépendance signifie que le "profil" de la densité en x ne change pas quelle que soit la valeur de y (et vice-versa). La surface de densité $z = f(x, y)$ peut être "séparée" en une fonction de x multipliée par une fonction de y.

La covariance se définit et se calcule de manière similaire.

Définition : Covariance (cas continu)

La **covariance** de X et Y mesure leur variation linéaire commune :

$$\mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y) \, \mathrm{d}x \, \mathrm{d}y$$

La formule de calcul reste la même.

Théorème : Formule de calcul de la Covariance

Une formule plus simple pour le calcul de la covariance est :

$$\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$$

où $E[XY]$ est calculé via LOTUS : $E[XY] = \iint xy f(x, y) \, \mathrm{d}x\mathrm{d}y$.

La preuve est identique à celle du cas discret.

Preuve

La preuve est identique à celle vue pour les variables discrètes, car elle ne repose que sur la

linéarité de l'espérance, qui est vraie aussi dans le cas continu. Soit $\mu_X = E[X]$ et $\mu_Y = E[Y]$.

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\
&= E[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y] \\
&= E[XY] - E[X\mu_Y] - E[Y\mu_X] + E[\mu_X\mu_Y] \\
&= E[XY] - \mu_Y E[X] - \mu_X E[Y] + \mu_X\mu_Y \\
&= E[XY] - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y \\
&= E[XY] - E[X]E[Y]
\end{aligned}$$

La relation entre indépendance et covariance reste la même.

Remarque : Indépendance et Covariance

Si X et Y sont indépendantes, alors $\mathrm{Cov}(X,Y) = 0$. Cependant, la réciproque n'est **pas** toujours vraie pour les variables aléatoires en général (bien qu'elle le soit dans certains cas importants comme pour les variables gaussiennes). Une covariance nulle signifie seulement une absence de *relation linéaire*, mais il peut exister d'autres formes de dépendance.

5 Espérance et Variance

5.1 Espérance d'une variable aléatoire discrète

Maintenant que nous avons défini les variables aléatoires discrètes et leur distribution (PMF), l'étape suivante est de résumer ces distributions. La mesure la plus importante est leur "centre", ou leur valeur moyenne.

Définition : Espérance (cas discret)

L'espérance (ou valeur attendue) d'une variable aléatoire discrète X, qui prend les valeurs distinctes $x_1, x_2, \ldots$, est définie par :

$$E(X) = \sum_j x_j P(X = x_j)$$

Cette formule est une moyenne pondérée de toutes les valeurs possibles.

Intuition

L'espérance représente la valeur moyenne que l'on obtiendrait si l'on répétait l'expérience un très grand nombre de fois. C'est le **centre de gravité** de la distribution de probabilité. Si les probabilités étaient des masses placées sur une tige aux positions x_j, l'espérance serait le point d'équilibre.

L'exemple le plus simple est le lancer d'un dé.

Exemple : Lancer d'un dé

Soit X le résultat d'un lancer de dé équilibré. Chaque face a une probabilité de $1/6$. L'espérance est :

$$E(X) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = \frac{21}{6} = 3.5$$

Même si 3.5 n'est pas un résultat possible, c'est la valeur moyenne sur un grand nombre de lancers.

5.2 Espérance d'une variable aléatoire continue

Lorsque la variable aléatoire X est continue, sa distribution est décrite par une fonction de densité de probabilité (PDF), $f(x)$. L'espérance est définie de manière analogue, en remplaçant la somme par une intégrale.

Définition : Espérance (cas continu)

L'espérance (ou valeur attendue) d'une variable aléatoire continue X avec une fonction de densité $f(x)$ est définie par :

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$$

L'intégrale doit être absolument convergente, c'est-à-dire $\int_{-\infty}^{\infty} |x| f(x)\, dx < \infty$.

Intuition

L'intuition du **centre de gravité** est toujours valable. Si la fonction de densité $f(x)$ représente la répartition de la masse sur une tige (l'axe des x), alors $E(X)$ est le point d'équilibre où la tige tiendrait en balance.

Exemple : Loi uniforme

Soit $X \sim \mathcal{U}(a, b)$. Sa densité est $f(x) = \frac{1}{b-a}$ pour $x \in [a, b]$, et 0 ailleurs.

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

L'espérance est le point milieu de l'intervalle, ce qui est intuitivement correct.

5.3 Linéarité de l'espérance

Le calcul de l'espérance deviendrait très fastidieux si nous devions toujours utiliser la définition. Heureusement, l'espérance possède une propriété fondamentale qui simplifie énormément les calculs.

Théorème : Linéarité de l'espérance

Pour toutes variables aléatoires X et Y (discrètes ou continues), et pour toute constante c, on a :

$$\begin{aligned} E(X+Y) &= E(X) + E(Y) \\ E(cX) &= cE(X) \end{aligned}$$

Cette propriété est extrêmement puissante car elle ne requiert pas que X et Y soient indépendantes.

La preuve de $E(cX) = cE(X)$ est directe à partir de la définition (discrète ou continue). La preuve pour la somme $E(X+Y)$ est plus complexe mais essentielle.

Preuve

La première propriété est directe.

- **Cas discret :** $E(cX) = \sum_x (cx)P(X=x) = c\sum_x xP(X=x) = cE(X)$
- **Cas continu :** $E(cX) = \int (cx)f(x)dx = c\int xf(x)dx = cE(X)$

Pour la seconde, $E(X+Y) = E(X) + E(Y)$, la preuve est analogue dans les deux cas.
Cas discret : Soit $S = X + Y$. L'espérance $E(S)$ se calcule en sommant sur toutes les paires possibles (x, y) avec la PMF jointe $P(X = x, Y = y)$:

$$\begin{aligned} E(X+Y) &= \sum_x \sum_y (x+y)P(X=x, Y=y) \\ &= \sum_x \sum_y xP(X=x, Y=y) + \sum_x \sum_y yP(X=x, Y=y) \\ &= \sum_x x\left(\sum_y P(X=x, Y=y)\right) + \sum_y y\left(\sum_x P(X=x, Y=y)\right) \end{aligned}$$

Par la loi des probabilités marginales, la somme interne $\sum_y P(X=x, Y=y)$ est $P(X=x)$, et de même $\sum_x P(X=x, Y=y) = P(Y=y)$.

$$E(X+Y) = \sum_x xP(X=x) + \sum_y yP(Y=y) = E(X) + E(Y)$$

Cas continu : La preuve est identique en remplaçant les sommes par des intégrales et la PMF

jointe par la PDF jointe $f(x, y)$:

$$\begin{aligned} E(X+Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf(x,y)\,dx\,dy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf(x,y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}x\left(\int_{-\infty}^{\infty}f(x,y)\,dy\right)dx + \int_{-\infty}^{\infty}y\left(\int_{-\infty}^{\infty}f(x,y)\,dx\right)dy \end{aligned}$$

Les intégrales internes sont les densités marginales $f_X(x) = \int f(x,y)dy$ et $f_Y(y) = \int f(x,y)dx$.

$$E(X+Y) = \int_{-\infty}^{\infty}xf_X(x)\,dx + \int_{-\infty}^{\infty}yf_Y(y)\,dy = E(X) + E(Y)$$

Notez que l'indépendance n'a jamais été requise pour cette preuve.

Cette propriété est incroyablement utile.

Intuition

Cette propriété formalise une idée très simple : "la moyenne d'une somme est la somme des moyennes". Si vous jouez à deux jeux de hasard, votre gain moyen total est simplement la somme de ce que vous gagnez en moyenne à chaque jeu, que les jeux soient liés ou non.

Cette propriété rend le calcul de l'espérance d'une somme trivial, comme le montre l'exemple des deux dés.

Exemple : Somme de deux dés

Soit X_1 le résultat du premier dé et X_2 celui du second. On sait que $E(X_1) = 3.5$ et $E(X_2) = 3.5$. Soit $S = X_1 + X_2$ la somme des deux dés. Grâce à la linéarité, on peut calculer l'espérance de la somme sans avoir à lister les 36 résultats possibles :

$$E(S) = E(X_1 + X_2) = E(X_1) + E(X_2) = 3.5 + 3.5 = 7$$

5.4 Espérance de la loi binomiale

Nous pouvons maintenant utiliser cette puissante propriété de linéarité pour trouver l'espérance de nos distributions de référence, en évitant des sommes complexes.

Théorème : Espérance de la loi binomiale

Si $X \sim \text{Bin}(n, p)$, alors son espérance est $E(X) = np$.

Ce résultat est profondément intuitif.

Intuition

Ce résultat est très naturel. Si vous lancez une pièce 100 fois ($n = 100$) avec une probabilité de 50% d'obtenir Pile ($p = 0.5$), vous vous attendez en moyenne à obtenir $100 \times 0.5 = 50$ Piles. La formule np généralise cette idée.

La preuve formelle est un exemple parfait de l'élégance de la linéarité, utilisant les variables indicatrices.

Preuve

Le calcul direct de l'espérance avec la PMF binomiale est possible, mais long. En utilisant la linéarité de l'espérance, on obtient une preuve beaucoup plus courte et élégante.
On peut voir une variable binomiale X comme la somme de n variables de Bernoulli indépendantes, $X = I_1 + I_2 + \cdots + I_n$, où chaque I_j représente le succès (1) ou l'échec (0) du j-ième essai.
Chaque I_j a pour espérance $E(I_j) = 1 \cdot p + 0 \cdot (1-p) = p$.

Par linéarité de l'espérance, on a :

$$E(X) = E(I_1) + E(I_2) + \cdots + E(I_n) = \underbrace{p + p + \cdots + p}_{n \text{ fois}} = np$$

5.5 Espérance de la loi géométrique

Calculons maintenant l'espérance pour la loi qui modélise le temps d'attente.

Théorème : Espérance de la loi géométrique

L'espérance d'une variable aléatoire $X \sim \text{Geom}(p)$ (comptant le nombre d'échecs) est :

$$E(X) = \frac{1-p}{p} = \frac{q}{p}$$

L'intuition est aussi très forte ici :

Intuition

Si un événement a 1 chance sur 10 de se produire ($p = 0.1$), il est logique de penser qu'il faudra en moyenne 9 échecs ($q/p = 0.9/0.1 = 9$) avant qu'il ne se produise. L'espérance du nombre total d'essais (échecs + 1 succès) serait alors $1/p$.

Contrairement à la loi binomiale, la preuve la plus directe ne repose pas sur la linéarité mais sur une manipulation de séries.

Preuve : Démonstration de l'espérance géométrique via les séries entières

Soit $X \sim \text{Geom}(p)$, où X compte le nombre d'échecs avant le premier succès. La PMF est $P(X = k) = q^k p$ pour $k = 0, 1, 2, \ldots$, avec $q = 1 - p$.
Par définition, l'espérance est :

$$E(X) = \sum_{k=0}^{\infty} k \cdot P(X = k) = \sum_{k=0}^{\infty} k q^k p$$

Le terme pour $k = 0$ est nul, on peut donc commencer la somme à $k = 1$:

$$E(X) = p \sum_{k=1}^{\infty} k q^k$$

L'astuce consiste à reconnaître que la somme ressemble à la dérivée d'une série géométrique. Rappelons la formule de la série géométrique pour $|q| < 1$:

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$$

En dérivant les deux côtés par rapport à q, on obtient :

$$\frac{d}{dq}\left(\sum_{k=0}^{\infty} q^k\right) = \frac{d}{dq}\left(\frac{1}{1-q}\right)$$

$$\sum_{k=1}^{\infty} k q^{k-1} = \frac{1}{(1-q)^2}$$

Pour faire apparaître ce terme dans notre formule d'espérance, on factorise q dans la somme :

$$E(X) = p \cdot q \sum_{k=1}^{\infty} k q^{k-1}$$

On peut maintenant remplacer la somme par son expression analytique :

$$E(X) = p \cdot q \cdot \frac{1}{(1-q)^2}$$

Puisque $p = 1 - q$, on a :

$$E(X) = p \cdot q \cdot \frac{1}{p^2} = \frac{q}{p}$$

Ce qui démontre que l'espérance du nombre d'échecs avant le premier succès est $\frac{q}{p}$.

5.6 Loi du statisticien inconscient (LOTUS)

Souvent, nous ne sommes pas intéressés par l'espérance de X elle-même, mais par l'espérance d'une fonction de X, par exemple $E(X^2)$ ou $E(e^X)$.

Théorème : Théorème de Transfert (LOTUS)

Si X est une variable aléatoire et $g(x)$ est une fonction de $\mathbb{R}$ dans $\mathbb{R}$, alors l'espérance de la variable aléatoire $g(X)$ est donnée par :

- **Cas discret :** $E[g(X)] = \sum_x g(x)P(X = x)$
- **Cas continu :** $E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)\,dx$

Ce théorème est utile car il évite d'avoir à trouver la distribution (PMF ou PDF) de $g(X)$.

La preuve dans le cas discret consiste simplement à regrouper les termes.

Preuve

Nous montrons la preuve pour le cas discret. La preuve pour le cas continu est plus technique (utilisant un changement de variable) et est omise.
Soit $Y = g(X)$. Par définition, l'espérance de Y est $E(Y) = \sum_y yP(Y = y)$. L'ensemble des valeurs y que Y peut prendre est $\{g(x) \mid x \in \text{support de } X\}$. Pour une valeur y donnée, l'événement $\{Y = y\}$ est l'union de tous les événements $\{X = x\}$ tels que $g(x) = y$.

$$P(Y = y) = P(g(X) = y) = \sum_{x:g(x)=y} P(X = x)$$

En substituant cela dans la définition de $E(Y)$:

$$E(Y) = \sum_y y \left(\sum_{x:g(x)=y} P(X = x) \right)$$

On peut réécrire y comme $g(x)$ à l'intérieur de la seconde somme :

$$E(g(X)) = \sum_y \sum_{x:g(x)=y} g(x)P(X = x)$$

Cette double somme parcourt toutes les valeurs de y, et pour chaque y, elle parcourt tous les x correspondants. Cela revient à simplement sommer sur tous les x possibles dès le départ :

$$E[g(X)] = \sum_x g(x)P(X = x)$$

Ce théorème justifie son nom : c'est ce que l'on ferait "inconsciemment".

Intuition

Pour trouver la valeur moyenne d'une fonction d'une variable aléatoire (par exemple, le carré du résultat d'un dé), vous n'avez pas besoin de déterminer d'abord la distribution de ce carré. Vous pouvez simplement prendre chaque valeur possible du résultat original, lui appliquer la fonction, et pondérer ce nouveau résultat par la probabilité (ou densité) du résultat original.

Utilisons ce théorème pour calculer $E(X^2)$ pour notre dé.

Exemple : Calcul de $E(X^2)$ pour un dé (discret)

Soit X le résultat d'un lancer de dé. Calculons l'espérance de $Y = X^2$. La fonction est $g(x) = x^2$.

$$\begin{aligned} E(X^2) &= \sum_{k=1}^{6} k^2 P(X = k) \\ &= 1^2\left(\frac{1}{6}\right) + 2^2\left(\frac{1}{6}\right) + 3^2\left(\frac{1}{6}\right) + 4^2\left(\frac{1}{6}\right) + 5^2\left(\frac{1}{6}\right) + 6^2\left(\frac{1}{6}\right) \\ &= \frac{1+4+9+16+25+36}{6} = \frac{91}{6} \approx 15.17 \end{aligned}$$

Exemple : Calcul de $E(X^2)$ pour une loi uniforme (continu)

Soit $X \sim \mathcal{U}(0,1)$. Sa densité est $f(x) = 1$ sur $[0,1]$. Calculons l'espérance de $Y = X^2$. La fonction est $g(x) = x^2$.

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} g(x) f(x)\, dx = \int_0^1 x^2 \cdot 1\, dx \\ &= \left[\frac{x^3}{3}\right]_0^1 = \frac{1}{3} \end{aligned}$$

5.7 Variance

L'espérance nous donne le centre d'une distribution, mais elle ne dit rien sur sa "largeur" ou sa "dispersion". C'est le rôle de la variance.

Définition : Variance et écart-type

La **variance** d'une variable aléatoire X mesure la dispersion de sa distribution autour de son espérance $\mu = E(X)$. Elle est définie par :

$$\mathrm{Var}(X) = E\left[(X-\mu)^2\right]$$

Concrètement, cela se traduit par (en utilisant LOTUS avec $g(x) = (x-\mu)^2$) :

- **Cas discret :** $\mathrm{Var}(X) = \sum_x (x-\mu)^2 P(X = x)$
- **Cas continu :** $\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$

La racine carrée de la variance est appelée l' **écart-type** :

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$$

L'idée est de mesurer l'écart quadratique moyen à l'espérance.

Intuition

La variance est la "distance carrée moyenne à la moyenne". On prend l'écart de chaque valeur par rapport à la moyenne, on le met au carré (pour que les écarts positifs et négatifs ne s'annulent pas), puis on en calcule la moyenne. L'écart-type est souvent plus interprétable car il ramène cette mesure de dispersion dans les mêmes unités que la variable aléatoire elle-même.

La définition $E[(X-\mu)^2]$ est excellente pour l'interprétation, mais pénible pour le calcul. Une formule alternative est presque toujours utilisée.

Théorème : Formule de calcul de la variance

Pour toute variable aléatoire X (discrète ou continue), une formule plus pratique pour le calcul de la variance est :

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2$$

La preuve est une simple expansion algébrique utilisant la linéarité de l'espérance (qui s'applique aux cas discrets et continus).

Preuve

Soit $\mu = E(X)$. On part de la définition de la variance :

$$\begin{aligned}
\mathrm{Var}(X) &= E[(X-\mu)^2] \\
&= E[X^2 - 2X\mu + \mu^2] \quad \text{(On développe le carré)} \\
&= E(X^2) - E(2\mu X) + E(\mu^2) \quad \text{(Par linéarité de l'espérance)} \\
&= E(X^2) - 2\mu E(X) + \mu^2 \quad \text{(Car } 2\mu \text{ et } \mu^2 \text{ sont des constantes)} \\
&= E(X^2) - 2\mu(\mu) + \mu^2 \quad \text{(Car } E(X) = \mu) \\
&= E(X^2) - 2\mu^2 + \mu^2 \\
&= E(X^2) - \mu^2 = E(X^2) - [E(X)]^2
\end{aligned}$$

Nous pouvons maintenant calculer la variance de notre lancer de dé.

Exemple : Variance d'un lancer de dé

Nous avons déjà calculé pour un dé que $E(X) = 3.5$ et $E(X^2) = 91/6$. On peut maintenant trouver la variance facilement :

$$\begin{aligned}
\mathrm{Var}(X) = E(X^2) - [E(X)]^2 &= \frac{91}{6} - (3.5)^2 \\
&= \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{91}{6} - \frac{49}{4} \\
&= \frac{182}{12} - \frac{147}{12} = \frac{35}{12} \approx 2.917
\end{aligned}$$

L'écart-type est $\mathrm{SD}(X) = \sqrt{35/12} \approx 1.708$.

Exemple : Variance de la loi uniforme

Calculons la variance de $X \sim \mathcal{U}(a,b)$. Nous avons trouvé $E(X) = \frac{a+b}{2}$. Nous devons d'abord calculer $E(X^2)$ en utilisant LOTUS :

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

On utilise maintenant la formule de calcul $\mathrm{Var}(X) = E(X^2) - [E(X)]^2$:

$$\begin{aligned}
\mathrm{Var}(X) &= \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 \\
&= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} \\
&= \frac{4(a^2+ab+b^2) - 3(a^2+2ab+b^2)}{12} \\
&= \frac{4a^2+4ab+4b^2-3a^2-6ab-3b^2}{12} \\
&= \frac{a^2-2ab+b^2}{12} = \frac{(b-a)^2}{12}
\end{aligned}$$

6 Distributions Multivariées et Concepts Associés

6.1 Distributions Jointes et Marginales

Jusqu'à présent, nous avons étudié les variables aléatoires isolément. Nous allons maintenant examiner comment analyser les relations entre *plusieurs* variables aléatoires.

Définition : Distribution Jointe (Cas Discret)

Pour deux variables aléatoires discrètes X et Y, la **distribution jointe** (ou loi jointe) spécifie la probabilité de chaque paire d'issues. La fonction de masse de probabilité jointe (joint PMF) est :

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

Si X prend ses valeurs dans un ensemble S et Y dans un ensemble T, alors la somme de toutes les probabilités jointes est égale à 1 :

$$\sum_{x \in S} \sum_{y \in T} P(X = x, Y = y) = 1$$

Cette loi jointe est la "carte" complète de toutes les issues possibles.

Intuition

La distribution jointe est la "carte" complète de toutes les issues possibles. Elle répond à la question : "Quelle est la probabilité que X prenne cette valeur ET que Y prenne cette autre valeur en même temps ?". Si vous imaginez un tableau à double entrée pour X et Y, la loi jointe est l'ensemble de toutes les probabilités à l'intérieur du tableau.

Cette "carte" complète contient toutes les informations. Si nous ne nous intéressons qu'à une seule variable, nous pouvons la "réduire" en calculant sa distribution marginale.

Définition : Distribution Marginale

À partir de la distribution jointe, on peut obtenir la distribution **marginale** (ou loi marginale) de chaque variable. Pour obtenir la probabilité que X prenne une valeur x, on somme sur toutes les valeurs possibles de Y :

$$P(X = x) = \sum_{y \in T} P(X = x, Y = y)$$

Visuellement, cela correspond à "écraser" le tableau de probabilités sur un seul de ses axes.

Intuition

Les distributions marginales sont les "ombres" ou "projections" de la carte jointe sur un seul axe. Si la loi jointe est un tableau, les lois marginales sont les totaux de chaque ligne et de chaque colonne, que l'on écrirait "dans la marge" du tableau. Elles nous disent la probabilité d'une issue pour X sans se soucier de ce qu'il advient de Y.

L'exemple le plus simple est le lancer de deux dés.

Exemple : Lois jointe et marginale

On lance un dé rouge (X) et un dé bleu (Y). Il y a 36 issues, chacune avec une probabilité de $1/36$. **Loi jointe** : $P(X = x, Y = y) = 1/36$ pour tout $x, y \in \{1, \ldots, 6\}$. Par exemple, $P(X = 2, Y = 5) = 1/36$.
Loi marginale de X : Cherchons $P(X = 2)$. C'est la probabilité d'obtenir 2 sur le dé rouge, quel que soit le résultat du bleu.

$$P(X = 2) = \sum_{y=1}^{6} P(X = 2, Y = y)$$

$$P(X = 2) = P(X = 2, Y = 1) + \cdots + P(X = 2, Y = 6)$$

$$P(X = 2) = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{6}{36} = \frac{1}{6}$$

Ceci est bien la loi d'un seul dé.

6.2 Espérance d'une fonction de deux variables

Maintenant que nous avons la loi jointe (la carte des probabilités), nous pouvons l'utiliser pour calculer l'espérance de n'importe quelle fonction qui dépend des deux variables, $g(X, Y)$.

Définition : Espérance d'une fonction $g(X, Y)$

L'espérance d'une fonction $g(X, Y)$ de deux variables aléatoires discrètes X et Y est une généralisation du théorème de transfert (LOTUS) :

$$E[g(X, Y)] = \sum_{x \in S} \sum_{y \in T} g(x, y) P(X = x, Y = y)$$

C'est la moyenne de g, pondérée par les probabilités jointes.

Intuition

C'est la valeur moyenne attendue de la fonction g. Pour la calculer, on prend chaque résultat possible de $g(x, y)$, on le pondère par la probabilité que cette combinaison (x, y) se produise (donnée par la loi jointe), et on somme le tout.

Le cas le plus important de $g(X, Y)$ est la somme $X + Y$.

Exemple

Espérance de $E[X + Y]$ Avec nos deux dés, calculons l'espérance de la somme $S = X + Y$. La fonction est $g(X, Y) = X + Y$.

$$E[X + Y] = \sum_{x=1}^{6} \sum_{y=1}^{6} (x + y)\, P(X = x, Y = y)$$

$$E[X + Y] = \sum_{x=1}^{6} \sum_{y=1}^{6} (x + y)\, \frac{1}{36}$$

Plutôt que de faire ce long calcul, on peut utiliser la linéarité de l'espérance (qui est un cas particulier de ce théorème) :

$$E[X + Y] = E[X] + E[Y] = 3.5 + 3.5 = 7.$$

6.3 Covariance et Corrélation

La linéarité $E[X + Y] = E[X] + E[Y]$ est un outil puissant. Mais l'espérance ne nous dit rien sur la *relation* entre X et Y. Pour cela, nous introduisons la covariance.

Définition : Covariance

La **covariance** entre deux variables aléatoires X et Y, avec pour moyennes respectives μ_X et μ_Y, mesure la façon dont elles varient ensemble.

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

Elle mesure la direction de leur relation.

Intuition

La covariance est positive si les variables ont tendance à "bouger" dans la même direction (quand X est au-dessus de sa moyenne, Y a tendance à l'être aussi). Elle est négative si elles bougent en sens opposé (quand X est au-dessus de sa moyenne, Y a tendance à être en dessous). Si elle est nulle, il n'y a pas de tendance linéaire entre elles.

La définition $E[(X - \mu_X)(Y - \mu_Y)]$ est bonne pour l'intuition, mais difficile à calculer. Une formule alternative est presque toujours utilisée.

Théorème : Formule de calcul de la covariance
Une formule computationnelle plus simple pour la covariance est :

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$$

La preuve est une simple expansion algébrique.

Preuve
Soit $\mu_X = E[X]$ et $\mu_Y = E[Y]$. On part de la définition :

$$\begin{aligned}
\text{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\
&= E[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y] \quad \text{(On développe)} \\
&= E[XY] - E[X\mu_Y] - E[Y\mu_X] + E[\mu_X\mu_Y] \quad \text{(Par linéarité)} \\
&= E[XY] - \mu_Y E[X] - \mu_X E[Y] + \mu_X\mu_Y \quad \text{(Les moyennes sont des constantes)} \\
&= E[XY] - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y \\
&= E[XY] - \mu_X\mu_Y \\
&= E[XY] - E[X]E[Y]
\end{aligned}$$

Voyons cette formule en action.

Exemple : Calcul de covariance
Cas 1 : Dés indépendants. X et Y sont les résultats de deux dés. $E[X] = 3.5, E[Y] = 3.5$. Calculons $E[XY]$. Puisqu'ils sont indépendants, $E[XY] = E[X]E[Y] = 3.5 \times 3.5 = 12.25$. $\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = 12.25 - 12.25 = 0$. La covariance est nulle, ce qui est attendu pour des variables indépendantes.
Cas 2 : Variables dépendantes. Soit X un lancer de dé, et $Y = 2X$. $E[X] = 3.5$. $E[Y] = E[2X] = 2E[X] = 7$. $E[XY] = E[X \cdot 2X] = E[2X^2] = 2E[X^2]$. On sait que $E[X^2] = \frac{1^2+\ldots+6^2}{6} = 91/6$. $E[XY] = 2(91/6) = 91/3$. $\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = \frac{91}{3} - (3.5)(7) = \frac{91}{3} - 24.5 = 30.33\ldots - 24.5 \approx 5.833$. La covariance est positive, ce qui est logique : si X est grand, Y l'est aussi.

La covariance est un bon indicateur de la direction de la relation, mais sa magnitude est difficile à interpréter. Pour cela, nous la normalisons.

Définition : Corrélation
La **corrélation** (ou coefficient de corrélation de Pearson, r) est une version normalisée de la covariance, qui se situe toujours entre -1 et 1.

$$\text{Corr}(X, Y) = r = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y}$$

La corrélation résout le problème des unités.

Intuition
Le problème de la covariance est qu'elle dépend des unités de X et Y (par ex., kg $\cdot$ cm). Si vous changez les unités (grammes et mètres), la valeur de la covariance change, même si la relation est identique. La corrélation résout ce problème : elle est sans unité. Un coefficient de +1 indique une relation linéaire positive parfaite, -1 une relation linéaire négative parfaite, et 0 une absence de relation linéaire.

Cette normalisation se comprend mieux en voyant la corrélation comme une covariance de variables standardisées.

Intuition : Interprétation de la formule
On peut voir la corrélation de Pearson comme un processus en 3 étapes :

1. **Centrer les variables :** On calcule l'écart de chaque valeur à sa moyenne ($x_i - \bar{x}$ et $y_i - \bar{y}$). Cela élimine "l'effet de base" (ex : une personne de 180cm vs 170cm ; la moyenne change mais les écarts relatifs restent les mêmes).
2. **Normaliser les variables :** On divise chaque écart par l'écart-type de sa variable ($z_{xi} =$

$(x_i - \bar{x})/\sigma_X$ et $z_{yi} = (y_i - \bar{y})/\sigma_Y$). Ces nouvelles variables Z_X et Z_Y sont **standardisées** : elles ont une moyenne de 0, un écart-type de 1, et sont sans unité.

3. **Calculer la covariance des variables standardisées :** La corrélation n'est rien d'autre que la covariance de ces deux nouvelles variables standardisées : $r = \text{Cov}(Z_X, Z_Y)$.

Parce que les deux variables sont maintenant sur la même échelle (écart-type de 1), leur covariance (la corrélation) ne peut pas dépasser 1 en valeur absolue.

Reprenons notre exemple de dépendance parfaite :

Exemple : Calcul de corrélation

Reprenons l'exemple $Y = 2X$, où X est un lancer de dé. On a $\text{Cov}(X, Y) = 5.833... = 35/6$. $\text{Var}(X) = E[X^2] - E[X]^2 = 91/6 - (3.5)^2 = 35/12$. $\text{Var}(Y) = \text{Var}(2\text{X}) = 2^2\text{Var}(X) = 4(35/12) = 35/3$. $\sigma_X\sigma_Y = \sqrt{35/12} \cdot \sqrt{35/3} = \sqrt{(35 \cdot 35)/(12 \cdot 3)} = \sqrt{35^2/36} = 35/6$.

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y} = \frac{35/6}{35/6} = 1$$

La corrélation est de 1, ce qui est parfait : Y est une fonction linéaire parfaite de X.

6.4 Linéarité de la Covariance

Tout comme l'espérance, la covariance possède d'importantes propriétés de linéarité qui simplifient les calculs.

Définition : Linéarité de la Covariance

Pour des variables aléatoires X, Y, Z et des constantes a, b, c :

$$\text{Cov}(aX + bY + c, Z) = a\text{Cov}(X, Z) + b\text{Cov}(Y, Z)$$
$$\text{Cov}(X, aY + bZ + c) = a\text{Cov}(X, Y) + b\text{Cov}(X, Z)$$

La covariance est linéaire pour chaque argument (elle est **bilinéaire**). Les constantes additives disparaissent.

6.5 Résultats sur la Corrélation

La propriété la plus importante de la corrélation, qui découle de sa normalisation, est qu'elle est bornée.

Théorème : Bornes du Coefficient de Corrélation de Pearson

Pour toutes variables aléatoires X et Y, le coefficient de corrélation $\text{Corr}(X, Y)$ est borné :

$$-1 \leq \text{Corr}(X, Y) \leq 1$$

De plus, si $\text{Corr}(X, Y) = \pm 1$, alors il existe des constantes a et b telles que $Y = aX + b$, indiquant une relation linéaire parfaite.

La preuve de ces bornes repose sur le fait que la variance est toujours positive.

Preuve : Démonstration des Bornes de la Corrélation

La preuve repose sur le fait que la variance d'une variable aléatoire est toujours positive ou nulle.

Étape 1 : Variables Standardisées On définit les versions standardisées de X et Y :

$$X^* = \frac{X - \mu_X}{\sigma_X} \quad ; \quad Y^* = \frac{Y - \mu_Y}{\sigma_Y}$$

Par construction, $E[X^*] = E[Y^*] = 0$ et $\text{Var}(X^*) = \text{Var}(Y^*) = 1$.

Étape 2 : Covariance des variables standardisées Calculons la covariance de X^* et Y^*,

qui est, par définition, la corrélation de X et Y.

$$\begin{aligned}\text{Cov}(X^*, Y^*) &= \text{Cov}\left(\frac{X-\mu_X}{\sigma_X}, \frac{Y-\mu_Y}{\sigma_Y}\right) \\ &= \frac{1}{\sigma_X\sigma_Y}\text{Cov}(X-\mu_X, Y-\mu_Y) \\ &= \frac{1}{\sigma_X\sigma_Y}\text{Cov}(X,Y) \\ &= \text{Corr}(X,Y)\end{aligned}$$

Étape 3 : Variance de la somme et de la différence Considérons la variance de la somme et de la différence de ces variables standardisées.

$$\text{Var}(X^* + Y^*) = \text{Var}(X^*) + \text{Var}(Y^*) + 2\text{Cov}(X^*, Y^*)$$

$$\text{Var}(X^* + Y^*) = 1 + 1 + 2\text{Corr}(X,Y) = 2 + 2\text{Corr}(X,Y)$$

De même :

$$\text{Var}(X^* - Y^*) = \text{Var}(X^*) + \text{Var}(Y^*) - 2\text{Cov}(X^*, Y^*)$$

$$\text{Var}(X^* - Y^*) = 1 + 1 - 2\text{Corr}(X,Y) = 2 - 2\text{Corr}(X,Y)$$

Étape 4 : La variance est toujours ≥ 0 La variance d'une variable aléatoire ne peut pas être négative.

$$\text{Var}(X^* + Y^*) \geq 0 \implies 2 + 2\text{Corr}(X,Y) \geq 0 \implies \text{Corr}(X,Y) \geq -1$$

$$\text{Var}(X^* - Y^*) \geq 0 \implies 2 - 2\text{Corr}(X,Y) \geq 0 \implies \text{Corr}(X,Y) \leq 1$$

Ceci nous donne le résultat final :

$$-1 \leq \text{Corr}(X,Y) \leq 1$$

6.6 Standardisation et Non-Corrélation

Le processus de 'standardisation' utilisé dans la preuve de la corrélation et dans l'intuition est un concept fondamental en soi.

Définition : Variable Centrée Réduite
Soit X une variable aléatoire avec :

- moyenne $\mu_X = E[X]$
- écart-type $\sigma_X = \sqrt{\text{Var}(X)} > 0$

On définit sa version **centrée réduite** (standardisée) Z par :

$$Z = \frac{X - \mu_X}{\sigma_X}$$

Alors, Z a les propriétés suivantes :

1. **Centrée (moyenne nulle)** :

$$\begin{aligned}E[Z] &= E\left[\frac{X-\mu_X}{\sigma_X}\right] \\ &= \frac{1}{\sigma_X}E[X-\mu_X] \quad \text{(par linéarité, } \sigma_X \text{ est une constante)} \\ &= \frac{1}{\sigma_X}(E[X] - E[\mu_X]) \\ &= \frac{1}{\sigma_X}(E[X] - \mu_X) \quad \text{(car } \mu_X \text{ est une constante)} \\ &= \frac{\mu_X - \mu_X}{\sigma_X} = 0\end{aligned}$$

2. **Réduite (écart-type égal à 1)** :

$$\begin{aligned}\mathrm{Var}(Z) &= \mathrm{Var}\left(\frac{X-\mu_X}{\sigma_X}\right)\\ &= \left(\frac{1}{\sigma_X}\right)^2 \mathrm{Var}(X-\mu_X) \quad (\text{propriété } \mathrm{Var}(aY) = a^2\,\mathrm{Var}(Y))\\ &= \frac{1}{\sigma_X^2}\,\mathrm{Var}(X) \quad (\text{propriété } \mathrm{Var}(Y+b) = \mathrm{Var}(Y))\\ &= \frac{1}{\sigma_X^2}\cdot\sigma_X^2 = 1\end{aligned}$$

L'écart-type est donc $\sigma_Z = \sqrt{\mathrm{Var}(Z)} = \sqrt{1} = 1$.

Cette transformation permet de comparer des variables sur des échelles différentes.

Intuition : Que signifie centrer-réduire ?

Standardiser une variable se fait en deux temps, comme le montre la formule $Z = \frac{X-\mu_X}{\sigma_X}$:

1. **Centrer ($X-\mu_X$)** : C'est la première étape. On soustrait la moyenne μ_X. Cela revient à "déplacer" la distribution pour que son centre de gravité (sa moyenne) soit maintenant à 0. On ne regarde plus les valeurs brutes X, mais leurs **écarts** par rapport à la moyenne. (Propriété 1 : $E[Z] = 0$)
2. **Réduire (.../σ_X)** : C'est la deuxième étape. On divise ces écarts par l'écart-type σ_X. Cela revient à changer d'unité de mesure. L'ancienne unité (kg, cm, points...) est remplacée par une nouvelle unité universelle : "le nombre d'écarts-types". (Propriété 2 : $\mathrm{Var}(Z) = 1$)

Au final, une variable Z avec une valeur de 1.5 signifie "cette observation est 1.5 écarts-types au-dessus de la moyenne de sa distribution d'origine", peu importe ce que X mesurait.

Intuition : Analogie simple

Imaginons 2 élèves :

- Alice a des notes entre 80 et 100 (moyenne 90, écart-type 5).
- Bob a des notes entre 0 et 20 (moyenne 10, écart-type 4).

Comparer leurs notes brutes n'a pas de sens. Mais si on les standardise, on peut se demander : "quand Alice est 1 écart-type au-dessus de sa moyenne (une note de 95), Bob est-il aussi 1 écart-type au-dessus de sa propre moyenne (une note de 14) ?". La standardisation permet cette comparaison.

Exemple : Centrer-réduire un dé

Pour un lancer de dé X, on a $\mu_X = 3.5$ et $\sigma_X = \sqrt{35/12} \approx 1.708$. Si on obtient $X = 6$: $Z = (6-3.5)/1.708 \approx 1.46$. Si on obtient $X = 1$: $Z = (1-3.5)/1.708 \approx -1.46$. Obtenir 6 est à 1.46 écarts-types au-dessus de la moyenne.

Maintenant, formalisons le concept d'une covariance nulle.

Définition : Variables Non Corréelées

On dit que deux variables aléatoires X et Y sont **non corrélées** si leur covariance est nulle :

$$\mathrm{Cov}(X,Y) = 0$$

Cela est équivalent à dire que $E[XY] = E[X]E[Y]$.

Il est crucial de ne pas confondre "non corrélées" et "indépendantes".

Intuition

"Non corrélées" signifie qu'il n'y a **pas de relation linéaire** entre les variables. C'est plus faible que l'indépendance. Si X et Y sont indépendantes, elles sont forcément non corrélées. Mais l'inverse n'est pas vrai : X et Y peuvent être non corrélées (Cov=0) mais quand même dépendantes (par exemple si $Y = X^2$ pour un X centré).

6.7 Variance d'une Somme de Variables Aléatoires

Nous pouvons maintenant combiner nos connaissances de la variance et de la covariance pour répondre à une question cruciale : quelle est la variance d'une somme de variables, $X + Y$?

Théorème : Formules pour la variance d'une somme de deux variables

Pour deux variables aléatoires X et Y :

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

La preuve découle de la définition de la variance et de la linéarité de l'espérance.

Preuve

Soit $\mu_X = E[X]$ et $\mu_Y = E[Y]$.

$$\begin{aligned}
\text{Var}(X + Y) &= E\left[((X + Y) - E[X + Y])^2\right] \\
&= E\left[((X + Y) - (\mu_X + \mu_Y))^2\right] \quad \text{(Par linéarité de E)} \\
&= E\left[((X - \mu_X) + (Y - \mu_Y))^2\right] \quad \text{(On regroupe les termes)}
\end{aligned}$$

Posons $A = (X - \mu_X)$ et $B = (Y - \mu_Y)$.

$$\begin{aligned}
&= E[(A + B)^2] = E[A^2 + 2AB + B^2] \\
&= E[A^2] + 2E[AB] + E[B^2] \quad \text{(Par linéarité de E)}
\end{aligned}$$

Or, par définition :

$$\begin{aligned}
E[A^2] &= E[(X - \mu_X)^2] = \text{Var}(X) \\
E[B^2] &= E[(Y - \mu_Y)^2] = \text{Var}(Y) \\
E[AB] &= E[(X - \mu_X)(Y - \mu_Y)] = \text{Cov}(X, Y)
\end{aligned}$$

Donc, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$

Cette formule est fondamentale en finance et en ingénierie.

Intuition

La "volatilité" (variance) d'une somme n'est pas juste la somme des volatilités. Il faut ajouter le terme d'interaction (covariance). Si $\text{Cov}(X, Y) > 0$ (elles bougent ensemble), la somme est **plus** volatile que la somme des parties. Si $\text{Cov}(X, Y) < 0$ (elles bougent en sens inverse), elles s'amortissent mutuellement. La somme est **moins** volatile. C'est le principe de la diversification en finance.

Cela mène à un corollaire très important lorsque la covariance est nulle.

Théorème : Cas Particulier : Variables Non Corréelées

Si X et Y sont non corrélées (Cov=0), la formule se simplifie :

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

Preuve

Cela découle directement du théorème précédent. Si X et Y sont non corrélées, alors $\text{Cov}(X, Y) = 0$. Le terme $2\text{Cov}(X, Y)$ dans la formule générale $\text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$ devient nul, laissant :

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

C'est le cas pour nos dés indépendants.

Exemple : Variance d'une somme de dés

Soit $S = X + Y$ la somme de deux dés indépendants. Puisqu'ils sont indépendants, ils sont non corrélés ($\text{Cov}(X, Y) = 0$). On sait $\text{Var}(X) = 35/12$ et $\text{Var}(Y) = 35/12$.

$$\text{Var}(S) = \text{Var}(X) + \text{Var}(Y) = \frac{35}{12} + \frac{35}{12} = \frac{70}{12} = \frac{35}{6} \approx 5.833$$

C'est bien plus simple que de calculer $E[S^2]$ et $E[S]$.

On peut généraliser cette formule à N variables.

Théorème : Variance d'une somme de N variables
La formule générale pour la somme de N variables aléatoires $X_1, \ldots, X_n$ est :

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j)$$

Preuve
On utilise la propriété $\text{Var}(S) = \text{Cov}(S, S)$ et la bilinéarité de la covariance. Soit $S = \sum_{i=1}^{n} X_i$.

$$\begin{aligned}\text{Var}(S) = \text{Cov}(S, S) &= \text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \text{Cov}(X_i, X_j) \quad \text{(Par bilinéarité)}\end{aligned}$$

On peut séparer cette double somme en deux parties : le cas où $i = j$ et le cas où $i \neq j$.

$$\text{Var}(S) = \sum_{i=1}^{n} \text{Cov}(X_i, X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j)$$

Puisque $\text{Cov}(X_i, X_i) = E[(X_i - \mu_i)(X_i - \mu_i)] = E[(X_i - \mu_i)^2] = \text{Var}(X_i)$, on obtient :

$$\text{Var}(S) = \sum_{i=1}^{n} \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j)$$

Cette formule est au cœur de la théorie moderne du portefeuille.

Intuition
La variance totale d'un système (comme un portefeuille d'actions) est la somme de toutes les variances individuelles ("risques propres") plus la somme de **toutes** les paires de covariances ("risques d'interaction"). Dans un grand portefeuille, le nombre de termes de covariance (environ n^2) est bien plus grand que le nombre de termes de variance (n), donc le risque total est dominé par la façon dont les actifs interagissent.

6.8 Théorème sur la somme de lois de Poisson

Terminons avec un théorème très utile qui combine les idées d'indépendance et de somme de variables aléatoires pour une distribution spécifique.

Théorème : La Somme de v.a. de Poisson Indépendantes est Poisson
Soit $X_1, \ldots, X_k$ une séquence de variables aléatoires de Poisson indépendantes, avec des paramètres respectifs $\lambda_1, \ldots, \lambda_k$.

$$X_i \sim \text{Poisson}(\lambda_i) \quad \text{pour } i = 1, \ldots, k$$

Alors leur somme $Y = X_1 + \cdots + X_k$ suit également une loi de Poisson, dont le paramètre est la somme des paramètres :

$$Y \sim \text{Poisson}(\lambda_1 + \cdots + \lambda_k)$$

La preuve pour $k = 2$ (qui se généralise par récurrence) utilise l'indépendance et la formule du binôme de Newton.

Preuve : Preuve pour la somme de deux v.a.

Soit $X \sim \text{Poisson}(\lambda_1)$ et $Y \sim \text{Poisson}(\lambda_2)$, indépendantes. Soit $S = X + Y$. Nous cherchons $P(S = k)$. Pour que $S = k$, il faut que $X = j$ et $Y = k - j$, pour toutes les valeurs possibles de j (de 0 à k).

$$P(S = k) = \sum_{j=0}^{k} P(X = j, Y = k - j)$$

Par indépendance, $P(X = j, Y = k - j) = P(X = j)P(Y = k - j)$.

$$\begin{aligned} P(S = k) &= \sum_{j=0}^{k} \left(\frac{e^{-\lambda_1} \lambda_1^j}{j!} \right) \left(\frac{e^{-\lambda_2} \lambda_2^{k-j}}{(k-j)!} \right) \\ &= e^{-(\lambda_1+\lambda_2)} \sum_{j=0}^{k} \frac{\lambda_1^j \lambda_2^{k-j}}{j!(k-j)!} \end{aligned}$$

On multiplie et on divise par $k!$ pour faire apparaître le coefficient binomial :

$$\begin{aligned} P(S = k) &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \sum_{j=0}^{k} \frac{k!}{j!(k-j)!} \lambda_1^j \lambda_2^{k-j} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \sum_{j=0}^{k} \binom{k}{j} \lambda_1^j \lambda_2^{k-j} \end{aligned}$$

La somme est l'expansion du binôme de Newton pour $(\lambda_1 + \lambda_2)^k$.

$$P(S = k) = \frac{e^{-(\lambda_1+\lambda_2)} (\lambda_1 + \lambda_2)^k}{k!}$$

C'est la PMF d'une loi Poisson$(\lambda_1 + \lambda_2)$.

Ce résultat est très intuitif :

Intuition

Si des événements rares se produisent indépendamment à des taux constants, le nombre total d'événements se produisant est aussi un événement rare se produisant au taux total. Si les emails arrivent à $\lambda_1 = 5$/heure et les appels à $\lambda_2 = 10$/heure, les "communications totales" arrivent simplement à $\lambda = 5 + 10 = 15$/heure.

Exemple : Centre d'appels

Un centre d'appels reçoit des appels "Ventes" selon $X_1 \sim$ Poisson(10 appels/heure) et des appels "Support" selon $X_2 \sim$ Poisson(15 appels/heure). Les deux types d'appels sont indépendants. Le nombre total d'appels $Y = X_1 + X_2$ suit une loi $Y \sim$ Poisson$(10 + 15 =$ 25 appels/heure). La probabilité de recevoir exactement 20 appels en une heure est :

$$P(Y = 20) = \frac{e^{-25} 25^{20}}{20!}$$

7 La Loi Normale (ou Gaussienne)

7.1 Introduction et Fonction de Densité (PDF)

Après les lois discrètes et les lois continues de base (Uniforme, Exponentielle), nous abordons la distribution la plus célèbre et la plus utilisée en probabilités et statistiques.

Définition : Loi Normale

Une variable aléatoire continue X suit une **loi normale** (ou loi de Gauss) de paramètres μ (l'espérance) et σ^2 (la variance), notée $X \sim \mathcal{N}(\mu, \sigma^2)$, si sa fonction de densité de probabilité (PDF) est donnée par :

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

pour tout $x \in (-\infty, \infty)$, où $\sigma > 0$.

Cette formule, bien qu'imposante, décrit une forme très familière : la courbe en cloche.

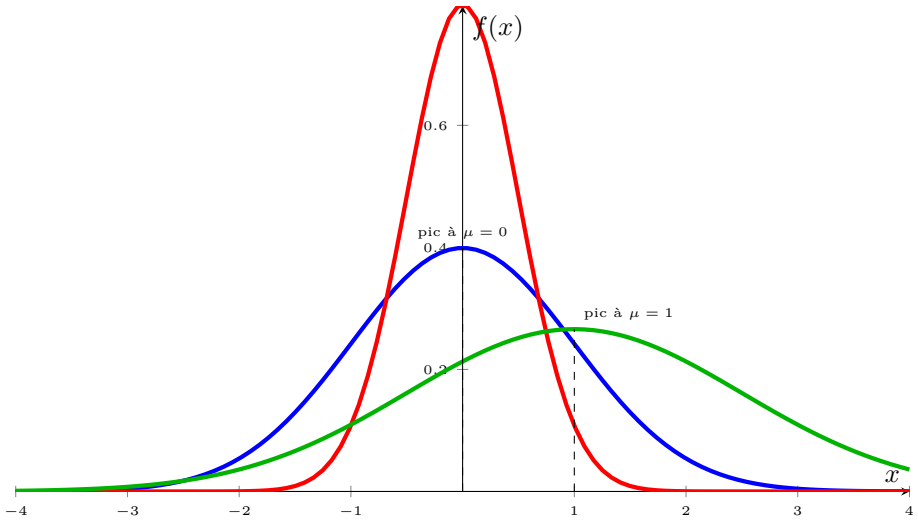
Intuition : La Courbe en Cloche

La loi normale est sans doute la distribution la plus importante en probabilités et statistiques. Pourquoi ? Parce qu'elle modélise remarquablement bien de nombreux phénomènes naturels et processus aléatoires où les valeurs tendent à se regrouper autour d'une moyenne, avec des écarts symétriques devenant de plus en plus rares à mesure qu'on s'éloigne de cette moyenne. Pensez à la taille des individus dans une population, aux erreurs de mesure répétées, ou même aux notes d'un grand groupe d'étudiants à un examen bien conçu.
Sa densité a une forme caractéristique de **cloche symétrique** :

- **Le Centre (μ)** : Le paramètre μ représente l'**espérance** (la moyenne) de la distribution. C'est le centre de symétrie de la courbe, là où la cloche atteint son **sommet**. C'est la valeur la plus probable (le mode) et aussi la valeur qui coupe la distribution en deux moitiés égales (la médiane). Changer μ *translate* la cloche horizontalement sans changer sa forme.
- **La Dispersion (σ)** : Le paramètre σ est l'**écart-type** (σ^2 est la variance). Il mesure la **dispersion** des valeurs autour de la moyenne μ. Géométriquement, σ contrôle la **largeur** de la cloche.
  - Un *petit* σ signifie que les données sont très concentrées autour de la moyenne, donnant une cloche **étroite et pointue**.
  - Un *grand* σ signifie que les données sont plus étalées, donnant une cloche **large et aplatie**.

  Les points d'inflexion de la courbe (là où la courbure change de sens) se situent exactement à $\mu \pm \sigma$.

La Courbe en Cloche (PDF de la Loi Normale)

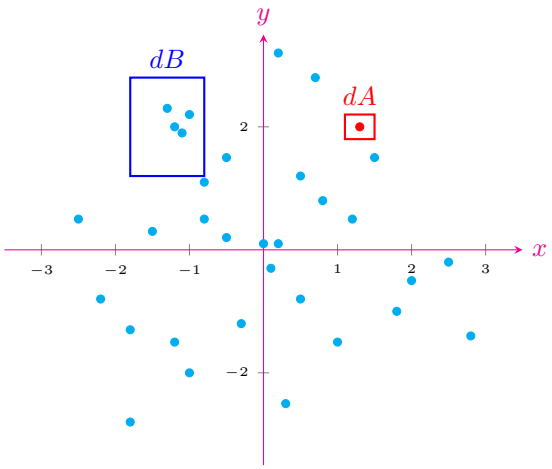


Influence de μ (position) et σ (largeur) sur la forme de la cloche.

Mais d'où vient cette formule spécifique ? Il existe une dérivation fascinante à partir d'hypothèses fondamentales sur les erreurs aléatoires (argument d'Herschel-Maxwell).

Preuve : Dérivation de la Densité Normale à partir des Principes Fondamentaux

Contexte Visuel : Imaginons un nuage de points dispersés autour d'une cible à l'origine $(0, 0)$, comme des impacts de fléchettes. Le graphique ci-dessous illustre cette dispersion. On s'intéresse à la probabilité de tomber dans une petite zone, comme dA, autour d'un point (x, y).



Objectif : Expliquer comment arriver à la formule mathématique de la courbe en cloche (densité de probabilité normale) en partant de principes fondamentaux sur les erreurs aléatoires.
1. Le Point de Départ : Densité et Aire dA Dans une distribution continue, la probabilité de tomber *exactement* sur un point (x, y) est nulle. On ne peut donc pas parler de "probabilité d'un point". On parle de la probabilité de tomber *dans une petite zone*, comme un rectangle $dA = dx \cdot dy$ autour du point (x, y). Cette probabilité, notée $P(\text{dans } dA)$, est *proportionnelle* à l'aire de la zone dA. La *constante de proportionnalité* est la **fonction de densité de probabilité** $p(x, y)$ évaluée en ce point. En d'autres termes, la densité $p(x, y)$ *représente* localement la concentration de probabilité. Ainsi, la probabilité de tomber dans la zone dA est approximativement :

$$P(\text{dans } dA) \approx p(x, y) \cdot dA$$

2. Les Hypothèses Fondamentales On pose deux hypothèses sur la nature de ces erreurs (représentées par la densité $p(x, y)$) :

1. **Indépendance des axes :** L'erreur horizontale (x) est indépendante de l'erreur verticale (y). Cela implique que la densité jointe $p(x, y)$ peut s'écrire comme le produit de la densité marginale sur x, notée $f(x)$, et de la densité marginale sur y, notée $f(y)$. Donc, $p(x, y) = f(x) \cdot f(y)$.
2. **Symétrie de rotation (Isotropie) :** La densité ne dépend que de la distance $r = \sqrt{x^2 + y^2}$ au centre, pas de l'angle. Il existe donc une fonction $\phi(r)$ telle que la densité en (x, y) est $p(x, y) = \phi(\sqrt{x^2 + y^2})$.

3. L'Équation Fonctionnelle En égalant les deux expressions pour la même densité $p(x, y)$ (à une constante près), on obtient :

$$f(x) \cdot f(y) = \phi(\sqrt{x^2 + y^2})$$

Pour $y = 0$, on a $f(x) \cdot f(0) = \phi(x)$. Posons $f(0) = \lambda$. Alors $\phi(x) = \lambda f(x)$. L'équation devient :

$$f(x) \cdot f(y) = \lambda f(\sqrt{x^2 + y^2})$$

4. Résolution de l'Équation Fonctionnelle Posons $g(x) = f(x)/\lambda$, avec $g(0) = 1$. L'équation se simplifie en :

$$g(x)g(y) = g(\sqrt{x^2 + y^2})$$

Posons $g(x) = h(x^2)$. L'équation devient $h(x^2)h(y^2) = h(x^2 + y^2)$. Avec $a = x^2$ et $b = y^2$, on a :

$$h(a)h(b) = h(a + b)$$

La solution continue de cette équation de Cauchy est $h(a) = e^{Aa}$ pour une constante A. Retour aux fonctions : $g(x) = h(x^2) = e^{Ax^2}$. $f(x) = \lambda g(x) = \lambda e^{Ax^2}$. Comme la densité doit diminuer loin du centre, A doit être négative. Posons $A = -k$ avec $k > 0$.

$$f(x) = \lambda e^{-kx^2}$$

5. Normalisation et Identification des Paramètres

1. **Condition** $\int f(x)dx = 1$: L'intégrale Gaussienne $\int_{-\infty}^{\infty} e^{-kx^2}\,\mathrm{d}x = \sqrt{\frac{\pi}{k}}$. Donc, $\int_{-\infty}^{\infty} f(x)dx = \lambda\sqrt{\frac{\pi}{k}} = 1 \implies \lambda = \sqrt{\frac{k}{\pi}}$.
2. **Lien avec la Variance (σ^2)** : Pour une distribution centrée, $\sigma^2 = E[X^2] = \int x^2 f(x)dx$.

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 \left(\sqrt{\frac{k}{\pi}} e^{-kx^2}\right) \mathrm{d}x = \sqrt{\frac{k}{\pi}} \left(\frac{1}{2k}\sqrt{\frac{\pi}{k}}\right) = \frac{1}{2k}$$

   Donc, $k = \frac{1}{2\sigma^2}$.
3. **Substitution Finale :** Remplaçons k dans λ et $f(x)$.

$$\lambda = \sqrt{\frac{1/(2\sigma^2)}{\pi}} = \frac{1}{\sigma\sqrt{2\pi}}$$

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

4. **Généralisation (Moyenne μ)** : Pour centrer la distribution sur μ, on remplace x par $(x - \mu)$ dans l'exposant :

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

C'est la fonction de densité de la loi normale $\mathcal{N}(\mu, \sigma^2)$.

7.2 La Loi Normale Centrée Réduite $\mathcal{N}(0,1)$

Avant d'explorer les propriétés de la loi normale générale, concentrons-nous sur son cas le plus simple et le plus fondamental.

Définition : Loi Normale Standard (ou Centrée Réduite)

Un cas particulier extraordinairement utile est la loi normale avec une moyenne $\mu = 0$ et une variance $\sigma^2 = 1$ (donc $\sigma = 1$). On l'appelle la **loi normale standard** ou **centrée réduite**, et on la note souvent Z. Sa PDF est traditionnellement notée $\phi(z)$:

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

Sa fonction de répartition (CDF), qui donne $P(Z \leq z)$, est notée $\Phi(z)$:

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \, \mathrm{d}t$$

Pourquoi cette version standard est-elle si importante ? Elle sert de référence universelle.

Intuition : La Référence Universelle et le Changement d'Unités

Pourquoi cette loi $\mathcal{N}(0,1)$ est-elle si centrale ? Imaginez que vous ayez des mesures en degrés Celsius ($\mathcal{N}(\mu_C, \sigma_C^2)$) et d'autres en degrés Fahrenheit ($\mathcal{N}(\mu_F, \sigma_F^2)$). Comment les comparer ? La loi normale standard fournit un **système d'unités universel**.
Toute variable normale $X \sim \mathcal{N}(\mu, \sigma^2)$ peut être transformée ("standardisée") en une variable $Z \sim \mathcal{N}(0,1)$ par un simple changement d'échelle et de position : $Z = (X - \mu)/\sigma$.
Cela signifie qu'au lieu de devoir calculer des aires (probabilités) pour une infinité de courbes en cloche différentes (une pour chaque paire μ, σ), on peut tout ramener à **une seule courbe de référence**, $\mathcal{N}(0,1)$. Les aires sous cette courbe standard ($\Phi(z)$) ont été calculées une fois pour toutes et sont disponibles dans des tables ou des logiciels. On n'a plus qu'à convertir notre problème dans cette "langue" standard, trouver la probabilité, et interpréter le résultat.

La notation est très standardisée pour cette loi.

Remarque : Notation ϕ et Φ

Les symboles ϕ (phi minuscule) pour la PDF et Φ (phi majuscule) pour la CDF de la loi normale standard sont quasi universels. Il est important de ne pas les confondre. $\phi(z)$ est la *hauteur* de la courbe en z, tandis que $\Phi(z)$ est l'*aire* sous la courbe à gauche de z.

Un détail technique important concerne le calcul de $\Phi(z)$.

Remarque : Absence de Primitive Simple

L'intégrale $\int e^{-t^2/2} \, \mathrm{d}t$, nécessaire pour calculer $\Phi(z)$, n'a **pas d'expression analytique** en termes de fonctions élémentaires (polynômes, exponentielles, log, sin, cos...). C'est une fonction spéciale, connue sous le nom de **fonction d'erreur** (liée à Φ par une transformation simple). C'est la raison pour laquelle on dépend de tables ou de calculs numériques pour obtenir les valeurs de $\Phi(z)$. Heureusement, ces outils sont omniprésents aujourd'hui.

7.3 Standardisation : Le Score Z

Formalisons cette transformation clé qui relie toute loi normale à la loi standard.

Théorème : Standardisation d'une Variable Normale

Si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors la variable Z définie par :

$$Z = \frac{X - \mu}{\sigma}$$

suit la loi normale standard, $Z \sim \mathcal{N}(0,1)$.

La preuve formelle utilise un changement de variable dans la fonction de répartition.

Preuve

Soit $F_X(x)$ la CDF de X et $F_Z(z)$ la CDF de Z. Nous voulons montrer que $F_Z(z) = \Phi(z)$.

$$\begin{aligned} F_Z(z) &= P(Z \leq z) \\ &= P\left(\frac{X-\mu}{\sigma} \leq z\right) \\ &= P(X - \mu \leq z\sigma) \\ &= P(X \leq \mu + z\sigma) \\ &= F_X(\mu + z\sigma) \end{aligned}$$

Par définition de la CDF de X :

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

Donc,

$$F_Z(z) = \int_{-\infty}^{\mu+z\sigma} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

Effectuons le changement de variable $u = (t-\mu)/\sigma$. Alors $t = \mu + u\sigma$ et $dt = \sigma du$. Les bornes d'intégration changent :

- Quand $t \to -\infty$, $u \to -\infty$.
- Quand $t = \mu + z\sigma$, $u = ((\mu + z\sigma) - \mu)/\sigma = z$.

L'intégrale devient :

$$F_Z(z) = \int_{-\infty}^{z} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}u^2} (\sigma du)$$

$$F_Z(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

C'est exactement la définition de $\Phi(z)$, la CDF de la loi normale standard. Ainsi, $Z \sim \mathcal{N}(0,1)$.

Cette transformation a une interprétation très concrète.

Intuition : Mesurer en "Unités d'Écart-Type"

Transformer X en Z s'appelle **standardiser** la variable. Le résultat, $z = \frac{x-\mu}{\sigma}$, est appelé le **Score Z** (ou cote Z). Ce score Z est une mesure *sans unité* qui indique **à combien d'écarts-types** une valeur observée x se situe par rapport à la moyenne μ de sa distribution.

- $z = 0$: x est exactement à la moyenne ($\mathbf{x} = \mu$).
- $z = +1$: x est un écart-type *au-dessus* de la moyenne ($\mathbf{x} = \mu + \sigma$).
- $z = -2$: x est deux écarts-types *en dessous* de la moyenne ($\mathbf{x} = \mu - \mathbf{2}\sigma$).

Cette transformation est extrêmement utile pour :

1. **Comparer des valeurs** issues de distributions normales différentes. Un score Z de $+1.5$ a toujours la même signification relative, que l'on parle de QI, de taille, ou de température.
2. **Calculer des probabilités** en utilisant la table unique de la loi $\mathcal{N}(0,1)$.

Un exemple classique est la comparaison de notes.

Exemple : Comparaison de Performances

Un étudiant A obtient 80 points à un examen où la moyenne est $\mu_A = 70$ et l'écart-type $\sigma_A = 5$. Un étudiant B obtient 85 points à un autre examen où $\mu_B = 75$ et $\sigma_B = 10$. Qui a le mieux réussi relativement à son groupe ?

Calculons les Z-scores :

$$Z_A = \frac{80-70}{5} = \frac{10}{5} = +2.0$$

$$Z_B = \frac{85-75}{10} = \frac{10}{10} = +1.0$$

L'étudiant A a un score Z plus élevé (+2.0 contre +1.0), ce qui signifie qu'il se situe plus d'écarts-types au-dessus de la moyenne de son groupe que l'étudiant B. L'étudiant A a donc relativement mieux réussi.

7.4 Propriétés Importantes de la Loi Normale

La loi normale possède des propriétés de stabilité remarquables sous certaines transformations.

Théorème : Stabilité par Transformation Linéaire

Si $X \sim \mathcal{N}(\mu, \sigma^2)$ et $Y = aX + b$ (avec $a \neq 0$), alors Y suit aussi une loi normale :

$$Y \sim \mathcal{N}(a\mu + b,\, (a\sigma)^2)$$

L'espérance est transformée linéairement ($E[aX+b] = aE[X]+b$), et la variance est multipliée par a^2 ($\text{Var}(aX + b) = a^2\text{Var}(X)$).

Preuve

Nous utilisons le fait que si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors $Z = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$. Exprimons X en fonction de Z : $X = \mu + \sigma Z$. Substituons cela dans l'expression de Y :

$$Y = a(\mu + \sigma Z) + b = (a\mu + b) + (a\sigma)Z$$

Posons $\mu_Y = a\mu + b$ et $\sigma_Y = |a|\sigma$. Alors $Y = \mu_Y + \sigma_Y Z$ (si $a > 0$) ou $Y = \mu_Y - \sigma_Y Z$ (si $a < 0$). Dans les deux cas, Y est une transformation linéaire d'une variable normale standard Z. La CDF de Y peut être exprimée en termes de la CDF Φ de Z. Si $a > 0$:

$$P(Y \leq y) = P(\mu_Y + a\sigma Z \leq y) = P(a\sigma Z \leq y - \mu_Y) = P\left(Z \leq \frac{y - \mu_Y}{a\sigma}\right) = \Phi\left(\frac{y - \mu_Y}{a\sigma}\right)$$

C'est la CDF d'une loi $\mathcal{N}(\mu_Y, (a\sigma)^2)$. Le cas $a < 0$ est similaire et mène au même résultat pour la distribution (la variance dépend de a^2). Ainsi, $Y \sim \mathcal{N}(a\mu + b, (a\sigma)^2)$.

Cette propriété est très utile pour les changements d'unités.

Exemple : Changement d'Unités

Si la température en Celsius T_C suit $\mathcal{N}(20, 5^2)$, quelle est la loi de la température en Fahrenheit $T_F = \frac{9}{5}T_C + 32$?
$a = 9/5$, $b = 32$.
Nouvelle moyenne : $E[T_F] = \frac{9}{5}(20) + 32 = 36 + 32 = 68$.
Nouvel écart-type : $\sigma_{T_F} = |a|\sigma_{T_C} = \frac{9}{5}(5) = 9$. Nouvelle variance : $\sigma^2_{T_F} = 9^2 = 81$.
Donc, $T_F \sim \mathcal{N}(68, 9^2)$.

Une autre propriété cruciale concerne la somme de variables normales indépendantes.

Théorème : Stabilité par Addition (Indépendance)

Si $X \sim \mathcal{N}(\mu_X, \sigma^2_X)$ et $Y \sim \mathcal{N}(\mu_Y, \sigma^2_Y)$ sont des variables aléatoires **indépendantes**, alors leur somme $S = X + Y$ suit aussi une loi normale :

$$S \sim \mathcal{N}(\mu_X + \mu_Y,\, \sigma^2_X + \sigma^2_Y)$$

Les moyennes s'ajoutent, et (grâce à l'indépendance) les variances s'ajoutent.

La preuve formelle de ce théorème est plus avancée et utilise généralement les fonctions caractéristiques ou les fonctions génératrices des moments.

Preuve : Idée de la preuve (via Fonctions Caractéristiques)

La fonction caractéristique $\varphi_X(t)$ d'une variable aléatoire X est définie comme $\varphi_X(t) = E[e^{itX}]$. Pour une loi normale $X \sim \mathcal{N}(\mu, \sigma^2)$, sa fonction caractéristique est $\varphi_X(t) = e^{i\mu t - \frac{1}{2}\sigma^2 t^2}$. Si X et Y sont indépendantes, la fonction caractéristique de leur somme $S = X + Y$ est le produit

de leurs fonctions caractéristiques : $\varphi_S(t) = \varphi_X(t)\varphi_Y(t)$.

$$\begin{aligned}\varphi_S(t) &= \left(e^{i\mu_X t - \frac{1}{2}\sigma_X^2 t^2}\right)\left(e^{i\mu_Y t - \frac{1}{2}\sigma_Y^2 t^2}\right) \\ &= e^{i(\mu_X+\mu_Y)t - \frac{1}{2}(\sigma_X^2+\sigma_Y^2)t^2}\end{aligned}$$

On reconnaît ici la fonction caractéristique d'une loi normale avec pour moyenne $\mu_X + \mu_Y$ et pour variance $\sigma_X^2 + \sigma_Y^2$. Comme la fonction caractéristique détermine de manière unique la distribution, on conclut que $S \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

Il est essentiel de se souvenir de la condition d'indépendance pour l'addition des variances.

Remarque : Attention à l'Indépendance

La propriété d'addition des variances ($\sigma_S^2 = \sigma_X^2 + \sigma_Y^2$) est cruciale et ne tient **que si** X **et** Y **sont indépendantes**. Si elles ne le sont pas, la variance de la somme inclut un terme de covariance : $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$. Cependant, la somme de variables normales (même dépendantes) reste normale (si elles sont conjointement normales).

Appliquons ce théorème à un exemple concret.

Exemple : Poids Total

Le poids d'une pomme suit $\mathcal{N}(150g, 10^2)$. Le poids d'une orange suit $\mathcal{N}(200g, 15^2)$. On suppose les poids indépendants. Quel est la loi du poids total d'une pomme et d'une orange ?
Soit P le poids de la pomme, O celui de l'orange. $T = P + O$.
$E[T] = E[P] + E[O] = 150 + 200 = 350g$.
$\text{Var}(T) = \text{Var}(P) + \text{Var}(O) = 10^2 + 15^2 = 100 + 225 = 325$.
Donc, $T \sim \mathcal{N}(350, 325)$. L'écart-type du poids total est $\sqrt{325} \approx 18.03g$.

7.5 La Règle Empirique (68-95-99.7)

Une conséquence directe des aires sous la courbe normale standard est une règle approximative très utile.

Théorème : Règle Empirique

Pour toute variable $X \sim \mathcal{N}(\mu, \sigma^2)$:

- $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.6827$ (Environ **68%** des valeurs dans $\mu \pm \sigma$).
- $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.9545$ (Environ **95%** des valeurs dans $\mu \pm 2\sigma$).
- $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.9973$ (Environ **99.7%** des valeurs dans $\mu \pm 3\sigma$).

Preuve : Dérivation à partir de $\Phi(z)$

Ces valeurs sont obtenues en calculant les aires sous la PDF de la loi normale standard $\mathcal{N}(0,1)$ entre les Z-scores correspondants.

- $P(-1 \leq Z \leq 1) = \Phi(1) - \Phi(-1) = \Phi(1) - (1 - \Phi(1)) = 2\Phi(1) - 1$. Avec $\Phi(1) \approx 0.8413$, on obtient $2(0.8413) - 1 \approx 0.6826$.
- $P(-2 \leq Z \leq 2) = \Phi(2) - \Phi(-2) = 2\Phi(2) - 1$. Avec $\Phi(2) \approx 0.9772$, on obtient $2(0.9772) - 1 \approx 0.9544$.
- $P(-3 \leq Z \leq 3) = \Phi(3) - \Phi(-3) = 2\Phi(3) - 1$. Avec $\Phi(3) \approx 0.99865$, on obtient $2(0.99865) - 1 \approx 0.9973$.

Ces valeurs sont souvent arrondies à 68

Cette règle fournit des repères très pratiques.

Intuition : Repères Essentiels sur la Cloche

Cette règle, dérivée directement des aires sous la courbe $\mathcal{N}(0,1)$ entre $z = \pm 1$, $z = \pm 2$ et $z = \pm 3$, fournit des repères extrêmement utiles pour interpréter l'écart-type σ. Elle nous dit où se trouve la grande majorité des données.
Une observation qui tombe en dehors de l'intervalle $\mu \pm 3\sigma$ est très inhabituelle (elle n'a que

0.3

7.6 Calcul de Probabilités Normales

En pratique, pour calculer une probabilité $P(a \leq X \leq b)$ pour une loi $\mathcal{N}(\mu, \sigma^2)$, on utilise systématiquement la standardisation.

Exemple : Utilisation du Z-score

Supposons que le QI d'une population suit $\mathcal{N}(100, 15^2)$. Quelle est la probabilité $P(X > 130)$? 1. **Standardiser :** $z = \frac{130-100}{15} = 2$. On cherche $P(Z > 2)$. 2. **Utiliser la CDF Standard :** $P(Z > 2) = 1 - P(Z \leq 2) = 1 - \Phi(2)$. 3. **Chercher dans la table / Calculer :** $\Phi(2) \approx 0.9772$. 4. **Résultat :** $P(X > 130) = 1 - 0.9772 = 0.0228$. Environ 2.3

Pour les intervalles, on utilise la propriété $P(a \leq Z \leq b) = \Phi(b) - \Phi(a)$.

Exemple : Probabilité entre deux valeurs

Quelle est la probabilité $P(85 \leq X \leq 115)$? $(\mu = 100, \sigma = 15)$
1. **Standardiser :** $z_1 = \frac{85-100}{15} = -1$, $z_2 = \frac{115-100}{15} = 1$. On cherche $P(-1 \leq Z \leq 1)$. 2. **Utiliser la CDF Standard :** $P(-1 \leq Z \leq 1) = \Phi(1) - \Phi(-1)$. 3. **Utiliser la symétrie :** $\Phi(-z) = 1 - \Phi(z)$. Donc $\Phi(-1) = 1 - \Phi(1)$. $P(-1 \leq Z \leq 1) = \Phi(1) - (1 - \Phi(1)) = 2\Phi(1) - 1$. 4. **Chercher dans la table / Calculer :** $\Phi(1) \approx 0.8413$. 5. **Résultat :** $P(85 \leq X \leq 115) \approx 2(0.8413) - 1 = 1.6826 - 1 = 0.6826$. (On retrouve la règle des 68

On peut aussi inverser le processus : trouver la valeur x correspondant à une probabilité donnée.

Exemple : Trouver une valeur pour une probabilité donnée (Problème Inverse)

Quel est le QI minimum requis pour être dans le top 10% de la population ? $(\mu = 100, \sigma = 15)$.
1. **Trouver le Z-score correspondant :** On cherche x tel que $P(X > x) = 0.10$. Cela équivaut à $P(Z > z) = 0.10$, où $z = (x - 100)/15$. Si $P(Z > z) = 0.10$, alors $P(Z \leq z) = \Phi(z) = 1 - 0.10 = 0.90$. 2. **Chercher dans la table inverse / Calculer :** On cherche la valeur z pour laquelle l'aire à gauche est 0.90 (le 90ème percentile). On trouve $z \approx 1.28$. 3. **Convertir en X :** On utilise la relation $z = (x - \mu)/\sigma$ pour trouver x : $1.28 = \frac{x-100}{15}$ $x = 100 + 1.28 \times 15 = 100 + 19.2 = 119.2$. Il faut un QI d'environ 119.2 pour être dans le top 10%.

8 Moments d'une distribution

8.1 Définitions fondamentales des moments

Après avoir défini l'espérance (μ) et la variance (σ^2), qui sont les moments d'ordre 1 et 2, nous pouvons généraliser cette idée pour capturer des informations plus subtiles sur la forme d'une distribution.

Définition : Types de Moments

Soit X une variable aléatoire ayant une espérance μ et une variance σ^2. Pour tout entier positif m, on définit les moments suivants :

- **m-ième moment (non centré)** : $E[X^m]$.
- **m-ième moment centré** : $E[(X-\mu)^m]$.
- **m-ième moment standardisé** : $E\left[\left(\frac{X-\mu}{\sigma}\right)^m\right]$.

Les moments centrés et standardisés permettent d'étudier les propriétés de la distribution indépendamment de sa position (μ) et de son échelle (σ).

8.2 Asymétrie (Skewness)

Le premier moment nous donne la tendance centrale. Le deuxième moment (la variance) nous donne la dispersion. Le troisième moment, lui, va nous renseigner sur la *symétrie* de la distribution.

Définition : Asymétrie (Skewness)

L'**asymétrie** (ou *skewness*) d'une variable aléatoire X de moyenne μ et d'écart-type σ est définie comme le **troisième moment standardisé** :

$$\text{Skew}(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right].$$

Intuition : Comprendre la Formule du Skewness

Pour une variable aléatoire X de moyenne μ et d'écart-type σ, le **skewness** est défini comme :

$$\text{Skew}(X) = \frac{E[(X-\mu)^3]}{\sigma^3}$$

Logique du numérateur : le moment centré d'ordre 3

- Le terme $(X-\mu)^3$ est le **cube de l'écart à la moyenne**
- Contrairement à $(X-\mu)^2$ (toujours positif), le cube **conserve le signe** de l'écart
- Il pondère différemment les observations à gauche et à droite de la moyenne

Interprétation intuitive

- **Skewness = 0 (Symétrique)** : La distribution est symétrique. Les écarts positifs et négatifs s'annulent. Typiquement : Moyenne = Médiane = Mode.
- **Skewness > 0 (Queue à droite)** : La distribution présente une queue longue à droite. Les grandes valeurs positives sont amplifiées par le cube. Les valeurs extrêmes tirent la moyenne vers la droite.
- **Skewness < 0 (Queue à gauche)** : La distribution présente une queue longue à gauche. Les écarts négatifs dominent. Les valeurs extrêmes tirent la moyenne vers la gauche.

Pourquoi σ^3 au dénominateur ?

- Le moment d'ordre 3 est homogène à des unités au cube
- On divise par σ^3 pour obtenir un coefficient **sans dimension**
- Permet la comparaison entre distributions de différentes échelles

Remarque : Pourquoi Standardiser ?

En standardisant d'abord ($\frac{X-\mu}{\sigma}$), la définition de Skew(X) ne dépend ni de la position (μ) ni de l'échelle (σ) de la distribution, ce qui est raisonnable puisque ces informations sont déjà fournies par la moyenne et l'écart-type. De plus, cette standardisation garantit que l'asymétrie est invariante par changement d'unité de mesure (par exemple, passer des pouces aux mètres n'affecte pas la valeur de l'asymétrie).

8.3 Propriétés de symétrie

Le skewness est une mesure numérique de l'asymétrie. Mais nous pouvons aussi définir la symétrie de manière formelle.

Définition : Symétrie d'une Variable Aléatoire

On dit qu'une variable aléatoire X a une distribution **symétrique** autour de μ si la variable $X - \mu$ a la même distribution que $\mu - X$. On dit aussi que X est symétrique ou que sa distribution est symétrique. Ces trois formulations ont le même sens.

Théorème : Symétrie en Termes de Fonction de Densité

Soit X une variable aléatoire continue de fonction de densité de probabilité (PDF) f. Alors, X est symétrique autour de μ si et seulement si :

$$f(x) = f(2\mu - x) \quad \text{pour tout } x.$$

Preuve : Preuve du Théorème de Symétrie

Soit F la fonction de répartition (CDF) de X. Si la symétrie tient, alors :

$$F(x) = P(X \leq x) = P(X - \mu \leq x - \mu) = P(\mu - X \leq x - \mu) = P(X \geq 2\mu - x) = 1 - F(2\mu - x).$$

En prenant la dérivée des deux côtés par rapport à x, on obtient :

$$f(x) = \frac{d}{dx}F(x) = \frac{d}{dx}[1 - F(2\mu - x)] = f(2\mu - x).$$

Cela démontre que la condition $f(x) = f(2\mu - x)$ est nécessaire et suffisante pour la symétrie.

8.4 Aplatissement (Kurtosis)

Après l'asymétrie (ordre 3), le moment d'ordre 4 nous informe sur "l'épaisseur" des queues de la distribution, c'est-à-dire la probabilité d'obtenir des valeurs très éloignées de la moyenne.

Définition : Kurtosis (Aplatissement)

Pour une variable aléatoire X de moyenne μ et d'écart-type σ, le **kurtosis** est défini comme le **quatrième moment standardisé** :

$$\text{Kurtosis}(X) = E\left[\left(\frac{X - \mu}{\sigma}\right)^4\right].$$

Dans la pratique, on utilise plus souvent le **kurtosis excessif** (ou excès de kurtosis), défini comme :

$$\text{Excess Kurtosis}(X) = E\left[\left(\frac{X - \mu}{\sigma}\right)^4\right] - 3.$$

La soustraction de 3 fait en sorte que le kurtosis d'une loi normale soit égal à 0.

Intuition : Comprendre la Kurtosis

Pour une variable aléatoire X, le **kurtosis** est défini comme :

$$\text{Kurt}(X) = \frac{E[(X - \mu)^4]}{\sigma^4}$$

et l'**excess kurtosis** (kurtosis excédentaire) comme : Excess Kurtosis = $\text{Kurt}(X) - 3$.

Pourquoi le moment d'ordre 4 ?

- Comme la variance, on utilise une puissance paire (pas d'effet de signe)
- La puissance 4 **amplifie énormément les écarts extrêmes**
- Mesure le **poids des queues** et la **concentration autour de la moyenne**

Interprétation intuitive (basée sur l'Excess Kurtosis)

- **Leptokurtique (Excess Kurtosis $>$ 0)** : Kurtosis total > 3. Distribution pointue avec des queues épaisses. Les événements extrêmes sont plus probables que pour une loi normale.
- **Mésocurtique (Excess Kurtosis $=$ 0)** : Kurtosis total $= 3$. C'est la référence (loi normale).
- **Platykurtique (Excess Kurtosis $<$ 0)** : Kurtosis total < 3. Distribution aplatie avec des queues légères et un centre large. Les événements extrêmes sont moins probables.

Application en finance

- Les rendements financiers ont souvent un excès de kurtosis positif
- Indique une probabilité plus élevée d'événements extrêmes que la loi normale
- Justifie le "vol smile" dans les options

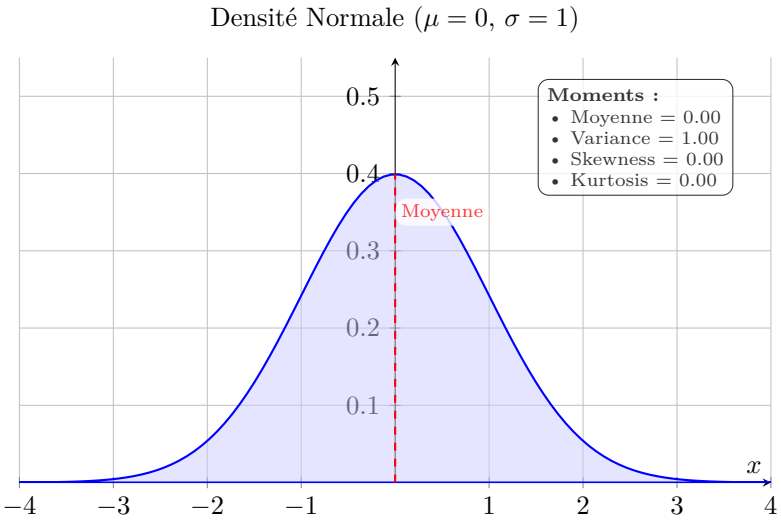
Pourquoi σ^4 au dénominateur ?

- Le moment d'ordre 4 est homogène à des unités^4
- On divise par σ^4 pour un coefficient **sans dimension**

8.5 Exemples de distributions

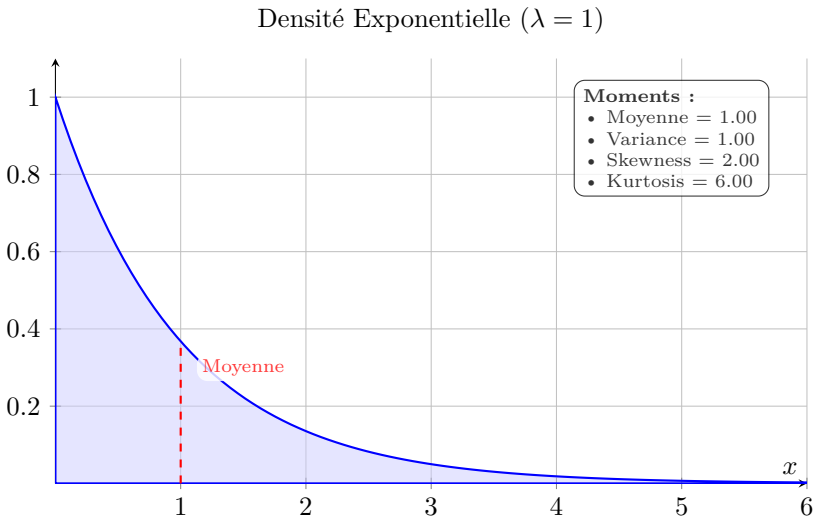
Pour bien fixer les idées, comparons le skewness et le kurtosis de plusieurs distributions classiques. Notez que dans les graphiques suivants, le "Kurtosis" affiché est l'*excess kurtosis* (centré à 0).

Exemple : La Distribution Normale (Mésokurtique)



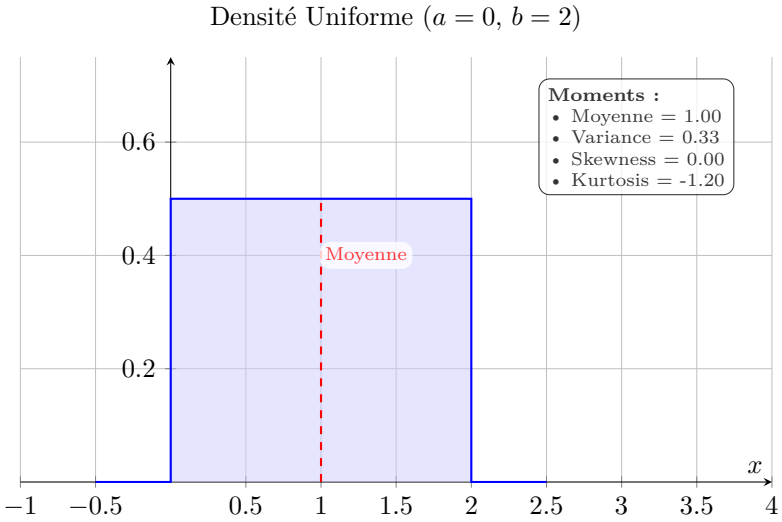
La distribution normale est l'archétype de la courbe en cloche. Imaginez une cible : la majorité des flèches touchent le centre, et plus on s'éloigne du centre, moins il y a de chances d'être touché. C'est une distribution parfaitement symétrique, ce qui se traduit par un **skewness nul (0.00)**. Son pic est ni trop pointu, ni trop plat : c'est notre point de référence, on dit qu'elle est **mésokurtique**, d'où son kurtosis de **0.00**. C'est la base de nombreuses analyses statistiques car elle modélise naturellement beaucoup de phénomènes.

Exemple : La Distribution Exponentielle (Asymétrique à Droite)



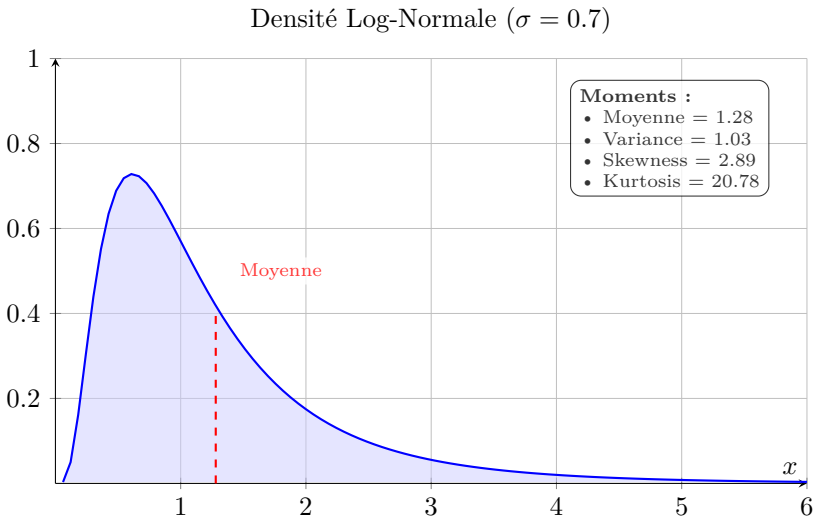
Imaginez le temps d'attente avant un événement rare, comme un appel téléphonique. La plupart du temps, l'appel arrive vite, mais il peut parfois y avoir de longues attentes. C'est exactement ce que modélise la distribution exponentielle : un pic à gauche et une longue queue à droite. Cela se traduit par un **skewness positif élevé (2.00)**, indiquant une asymétrie marquée. Elle est aussi **leptokurtique** (**kurtosis = 6.00**) : son pic est pointu, et la longue queue droite signifie qu'il y a une probabilité non négligeable de valeurs extrêmes.

Exemple : La Distribution Uniforme (Platykurtique)



La distribution uniforme, c'est le "tirage au sort parfait" : chaque valeur sur un intervalle a la même chance d'être tirée. Visuellement, c'est un rectangle, donc aucune valeur n'est privilégiée. Elle est symétrique (**skewness = 0.00**), mais contrairement à la normale, elle est "plate", sans pic central. Cela se traduit par un **kurtosis négatif (-1.20)**, ce qui signifie qu'elle est **platykurtique**. Elle est donc très différente des distributions avec un pic central comme la normale.

Exemple : La Distribution Log-Normale (Fortement Leptokurtique)



La log-normale est une distribution très asymétrique. Imaginez la richesse d'une population : la majorité est modeste, mais il existe une petite proportion de très riches, ce qui "étire" la droite de la courbe. Cela donne un **skewness très élevé (2.89)**. Elle est extrêmement **leptokurtique** (**kurtosis = 20.78**) : un pic très aigu et une queue droite très lourde. Cela signifie qu'il y a un risque élevé de valeurs extrêmement grandes, ce qui la rend très utile pour modéliser des phénomènes avec de rares événements extrêmes.

Nous avons défini les moments d'une *distribution* (moments de population), tels que $\mu = E[X]$ ou $\sigma^2 = E[(X - \mu)^2]$. Ce sont des valeurs théoriques, la "vérité" sous-jacente.

En pratique, nous ne connaissons presque jamais cette "vérité". Nous ne disposons que de données. Notre but est d'utiliser ces données pour *estimer* les moments de la population.

8.6 Moments d'échantillon (Sample Moments)

Définition : Moments d'Échantillon

Soit $X_1, X_2, \ldots, X_n$ un échantillon de n observations.

- La **moyenne d'échantillon** (notre "meilleure estimation" de μ) est :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- La **variance d'échantillon (non biaisée)** (notre "meilleure estimation" de σ^2) est :

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

De même, on peut calculer un *skewness d'échantillon* et un *kurtosis d'échantillon* en utilisant $\bar{X}$ et s, qui seront nos estimations du vrai skewness et du vrai kurtosis de la population.

Exemple : Application : Contrôle Qualité

Imaginez une usine qui produit des sacs de sucre de 1kg.

- **Population :** L'infinité de tous les sacs de sucre que la machine produira.
- **Moment de population (inconnu) :** Le poids moyen *réel* μ que la machine verse, et la variance *réelle* σ^2 (sa constance).
- **Problème :** Nous ne pouvons pas peser tous les sacs !
- **Solution :** Nous prélevons un **échantillon** de $n = 10$ sacs.
  Nous les pesons : $\{1002g, 998g, 1001g, 995g, 1003g, 1000g, 997g, 1005g, 999g, 1000g\}$.

- **Calcul des moments d'échantillon :**
  - $\bar{X} = (1002 + 998 + \cdots + 1000)/10 = 1000g$.
  - $s^2 = \frac{1}{10-1}\left((1002-1000)^2 + (998-1000)^2 + \dots\right) = 7.33g^2$.
- **Conclusion :** Notre meilleure estimation est que la machine est bien réglée sur $\mu = 1000g$. L'écart-type de notre échantillon est $s = \sqrt{7.33} \approx 2.7g$. Nous pouvons utiliser cela pour affirmer, par exemple, que 95% des sacs se situent probablement entre $1000 \pm 2s$ (si la distribution est normale).

Remarque : L'Intuition du "$n-1$"

Pourquoi diviser par $n-1$ pour la variance ? C'est la **correction de Bessel**.
Imaginez un échantillon de 1 seule personne ($n = 1$). Sa taille est 170cm.

- Quelle est la moyenne de l'échantillon ? $\bar{X} = 170$ cm.
- Quelle est la variance de l'échantillon ? $\sum(X_i - \bar{X})^2 = (170-170)^2 = 0$.
- Si on divisait par $n = 1$, on estimerait que la variance de la population est 0. C'est absurde ! Cela voudrait dire que tout le monde mesure 170cm.

En divisant par $n-1$ (donc $1-1=0$), la formule devient $0/0$ (indéfinie), ce qui nous dit à juste titre : "Je ne peux pas estimer la dispersion avec une seule personne."
Intuition plus générale : Nous "perdons un degré de liberté". Pour calculer la variance, nous avons besoin de connaître la moyenne. Mais nous ne connaissons pas la vraie moyenne μ. Nous devons donc utiliser $\bar{X}$, une *estimation*. Le fait d'utiliser une estimation calculée *à partir de ce même échantillon* introduit un léger biais (nos données sont, par définition, centrées sur $\bar{X}$). Diviser par $n-1$ au lieu de n "gonfle" légèrement le résultat pour compenser ce biais.

8.7 Fonctions génératrices des moments (MGF)

Définition : Fonction Génératrice des Moments (MGF)

La **fonction génératrice des moments** (MGF) d'une variable aléatoire X, notée $M_X(t)$, est définie comme :

$$M_X(t) = E[e^{tX}]$$

Intuition : L'ADN, le Code-Barres, ou le Fichier .zip

Ce concept est abstrait, alors utilisons des analogies :
Analogie 1 : L'ADN ou l'Empreinte Digitale

- La MGF est l'**empreinte digitale unique** d'une distribution.
- Elle "compresse" *toutes* les informations sur votre distribution (moyenne, variance, skewness, kurtosis, etc.) en une seule, unique fonction.
- Si deux distributions ont la même MGF, elles sont identiques. C'est la **propriété d'unicité**.

Analogie 2 : Le Code-Barres

- Pensez à une distribution (ex : Loi Normale) comme à un produit au supermarché.
- La MGF, $M_X(t)$, est son **code-barres unique**.
- Le processus de "génération de moments" (que nous verrons ci-dessous) est le **scanner**.
- En scannant le code-barres ($M_X(t)$), vous pouvez obtenir n'importe quelle information :
- Scan 1 ($M'_X(0)$) $\rightarrow$ vous donne le prix ($E[X]$).
- Scan 2 ($M''_X(0)$) $\rightarrow$ vous donne le poids ($E[X^2]$).
- Scan 3 ($M'''_X(0)$) $\rightarrow$ vous donne le pays d'origine ($E[X^3]$).

Pourquoi e^{tX} ? La "magie" vient du développement en série de Taylor de e^x :

$$e^{tX} = 1 + (tX) + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \dots$$

Quand on prend l'espérance, $E[\cdot]$, les puissances de X (c'est-à-dire $X, X^2, X^3 \dots$) apparaissent. Ce sont les moments ! La MGF "stocke" tous ces moments en les organisant comme coefficients d'un polynôme infini en t.

8.8 Génération des moments via les MGF

Théorème : Moments par Dérivation

Si la MGF $M_X(t)$ existe, alors le m-ième moment non centré $E[X^m]$ est la m-ième dérivée de $M_X(t)$, évaluée en $t = 0$:

$$E[X^m] = \frac{d^m}{dt^m} M_X(t)\Big|_{t=0} = M_X^{(m)}(0)$$

Exemple : Application : La Loi de Poisson

Une loi de Poisson modélise le nombre d'événements (ex : appels à un centre d'appels) par heure. Soit $X \sim \text{Poisson}(\lambda)$, où λ est le nombre moyen d'appels.
La MGF (l'ADN) d'une loi de Poisson est (on l'admet) :

$$M_X(t) = e^{\lambda(e^t - 1)}$$

Utilisons notre "scanner" (les dérivées) pour trouver les moments.
1. Trouver la Moyenne $E[X]$: On dérive une fois (règle de la chaîne) :

$$M'_X(t) = \frac{d}{dt}\left(e^{\lambda(e^t-1)}\right) = \underbrace{e^{\lambda(e^t-1)}}_{\text{répète}} \cdot \underbrace{(\lambda e^t)}_{\text{dérivée interne}}$$

Maintenant, on évalue en $t = 0$:

$$E[X] = M'_X(0) = e^{\lambda(e^0-1)} \cdot (\lambda e^0) = e^{\lambda(1-1)} \cdot (\lambda \cdot 1) = e^0 \cdot \lambda = 1 \cdot \lambda = \lambda$$

Résultat : La moyenne est λ, ce qui est la définition même du paramètre de la loi de Poisson. Parfait.
2. Trouver $E[X^2]$ (pour la variance) : On dérive une seconde fois (règle du produit sur $M'_X(t) = (\lambda e^t) \cdot (e^{\lambda(e^t-1)})$) :

$$M''_X(t) = \underbrace{(\lambda e^t)}_{\text{dérivée de u}} \cdot \underbrace{(e^{\lambda(e^t-1)})}_{\text{v}} + \underbrace{(\lambda e^t)}_{\text{u}} \cdot \underbrace{(e^{\lambda(e^t-1)} \cdot \lambda e^t)}_{\text{dérivée de v}}$$

Maintenant, on évalue en $t = 0$ (tous les e^0 deviennent 1) :

$$E[X^2] = M''_X(0) = (\lambda \cdot 1) \cdot (e^{\lambda(1-1)}) + (\lambda \cdot 1) \cdot (e^{\lambda(1-1)} \cdot \lambda \cdot 1)$$

$$E[X^2] = (\lambda) \cdot (e^0) + (\lambda) \cdot (e^0 \cdot \lambda) = \lambda \cdot 1 + \lambda \cdot (1 \cdot \lambda) = \lambda + \lambda^2$$

3. Trouver la Variance $\text{Var}(X)$: $\text{Var}(X) = E[X^2] - (E[X])^2 = (\lambda + \lambda^2) - (\lambda)^2 = \lambda$ **Résultat :** Nous avons prouvé par les MGF que pour une loi de Poisson, Moyenne = Variance = λ. C'est une propriété fondamentale de cette loi.

8.9 Sommes de variables aléatoires indépendantes via les MGF

C'est la super-puissance des MGF.

Théorème : MGF d'une Somme

Soient X et Y deux variables aléatoires **indépendantes**. Soit $S = X + Y$. Alors la MGF de S est le produit des MGF individuelles :

$$M_S(t) = M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$$

Intuition : La Magie de l'Exponentielle

Pourquoi est-ce vrai ? $M_{X+Y}(t) = E[e^{t(X+Y)}] = E[e^{tX} \cdot e^{tY}]$. Parce que X et Y sont indépendantes, $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$. Donc, $E[e^{tX} \cdot e^{tY}] = E[e^{tX}] \cdot E[e^{tY}] = M_X(t) \cdot M_Y(t)$. Les MGF transforment une opération analytiquement horrible (la "convolution" de densités) en une simple multiplication algébrique.

Exemple : Application : Portefeuille d'Actifs ou Tailles Humaines

C'est l'un des théorèmes les plus importants des statistiques. **Problème :** Soit X la taille d'un homme, $X \sim N(\mu_X, \sigma_X^2)$. Soit Y la taille d'une femme, $Y \sim N(\mu_Y, \sigma_Y^2)$. Si on les choisit au hasard, quelle est la loi de la somme de leurs tailles $S = X + Y$?

1. **ADN de X** : La MGF d'une loi Normale $N(\mu, \sigma^2)$ est $M(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$.
2. **ADN de X et Y** : $M_X(t) = \exp(\mu_X t + \frac{1}{2}\sigma_X^2 t^2)$ $M_Y(t) = \exp(\mu_Y t + \frac{1}{2}\sigma_Y^2 t^2)$
3. **ADN de $S = X + Y$** (on multiplie) : $M_S(t) = M_X(t) \cdot M_Y(t) = \exp(\mu_X t + \frac{1}{2}\sigma_X^2 t^2) \cdot \exp(\mu_Y t + \frac{1}{2}\sigma_Y^2 t^2)$
4. **Simplification** (en additionnant les exposants) : $M_S(t) = \exp\left((\mu_X t + \mu_Y t) + (\frac{1}{2}\sigma_X^2 t^2 + \frac{1}{2}\sigma_Y^2 t^2)\right)$ $M_S(t) = \exp\left((\mu_X + \mu_Y)t + \frac{1}{2}(\sigma_X^2 + \sigma_Y^2)t^2\right)$
5. **Conclusion (par Unicité)** : Regardez cet ADN ! C'est l'ADN d'une loi Normale ! Le nouveau μ est $(\mu_X + \mu_Y)$. La nouvelle σ^2 est $(\sigma_X^2 + \sigma_Y^2)$.

Résultat : Nous avons prouvé que **la somme de deux Normales indépendantes est une nouvelle Normale**. Si $X \sim N(175cm, 7^2)$ et $Y \sim N(165cm, 6^2)$, alors $S \sim N(340cm, 7^2+6^2 = 85)$. Notez que les écarts-types *ne s'additionnent pas* ($\sqrt{85} \approx 9.2 \neq 7+6$). Ce sont les variances qui s'additionnent.

9 Les Lois des Grands Nombres (LLN)

Dans la section précédente, nous avons fait une distinction cruciale entre les **moments de population** (les "vraies" valeurs théoriques, inconnues, comme μ et σ^2) et les **moments d'échantillon** (nos estimations calculées à partir des données, comme $\bar{X}$ et s^2).

Par exemple, nous avons défini la moyenne d'échantillon $\bar{X} = \frac{1}{n}\sum X_i$ comme notre "meilleure estimation" de la moyenne de population μ. Mais qu'est-ce qui nous garantit que cette estimation est "bonne" ? Qu'est-ce qui nous assure que si nous collections plus de données (en augmentant n), notre $\bar{X}$ se rapprocherait de μ ?

La réponse à cette question fondamentale est fournie par les **Lois des Grands Nombres (LLN)**. Elles forment le pont théorique entre les probabilités (la théorie) et les statistiques (la pratique).

Intuition : L'Idée Fondamentale : L'Exemple du Dé

Supposons que nous voulons connaître la valeur moyenne d'un lancer de dé équilibré.

- **Moment de Population :** Nous savons par la théorie que $\mu = E[X] = \frac{1+2+3+4+5+6}{6} = 3.5$.
- **Moments d'Échantillon :** Nous n'avons pas cette information, alors nous lançons le dé.
  - $n = 2$ lancers : On obtient (2, 6). $\bar{X}_2 = (2+6)/2 = 4.0$. (Assez loin de 3.5)
  - $n = 10$ lancers : On obtient (1, 6, 3, 3, 5, 2, 4, 1, 6, 4). $\bar{X}_{10} = 3.5$. (Pile dessus !)
  - $n = 100$ lancers : On obtiendra $\bar{X}_{100} \approx 3.48$ (par exemple).
  - $n = 1,000,000$ lancers : On obtiendra $\bar{X}_{1,000,000} \approx 3.5001$ (par exemple).

La Loi des Grands Nombres formalise cette intuition : à mesure que $n \to \infty$, la moyenne de notre échantillon $\bar{X}_n$ **converge** vers la vraie moyenne μ.
La distinction entre les lois "Faible" et "Forte" réside dans la *manière* dont nous définissons cette convergence.

9.1 L'Inégalité de Chebyshev

Avant de prouver la Loi Faible, nous avons besoin d'un outil fondamental qui relie la variance d'une variable à la probabilité qu'elle s'éloigne de sa moyenne. C'est l'Inégalité de Chebyshev.

Sa puissance réside dans son universalité : elle s'applique à *n'importe quelle* distribution, à condition qu'elle ait une moyenne et une variance finies.

Théorème : Inégalité de Chebyshev

Soit Y une variable aléatoire avec une espérance finie $\mu = E[Y]$ et une variance finie $\sigma^2 = \text{Var}(Y)$.
Alors, pour tout nombre réel $k > 0$:

$$P(|Y - \mu| \geq k) \leq \frac{\text{Var}(Y)}{k^2} = \frac{\sigma^2}{k^2}$$

Preuve : Preuve de l'Inégalité de Chebyshev

Nous présentons la preuve pour une variable aléatoire continue Y de densité $f(y)$. La preuve pour le cas discret est similaire en remplaçant les intégrales par des sommes.

1. Par définition, la variance σ^2 est $E[(Y-\mu)^2]$:

$$\sigma^2 = E[(Y-\mu)^2] = \int_{-\infty}^{\infty} (y-\mu)^2 f(y)dy$$

2. Nous pouvons scinder cette intégrale en deux parties : la région où Y est proche de μ ($|y-\mu| < k$) et la région où Y est loin de μ ($|y-\mu| \geq k$) :

$$\sigma^2 = \int_{|y-\mu|<k} (y-\mu)^2 f(y)dy + \int_{|y-\mu|\geq k} (y-\mu)^2 f(y)dy$$

3. L'intégrande $(y-\mu)^2 f(y)$ est toujours non-négative (un carré fois une densité). Par conséquent, la première intégrale est ≥ 0. En la supprimant, nous ne pouvons que diminuer la valeur totale :
$$\sigma^2 \geq \int_{|y-\mu|\geq k} (y-\mu)^2 f(y)dy$$
4. Maintenant, concentrons-nous sur la région d'intégration : $|y-\mu| \geq k$. Dans cette région, par définition, nous avons $(y-\mu)^2 \geq k^2$.
5. Nous pouvons remplacer $(y-\mu)^2$ par k^2 dans l'intégrale. Puisque nous remplaçons un terme par quelque chose de plus petit ou égal, la valeur de l'intégrale diminue (ou reste égale) :
$$\sigma^2 \geq \int_{|y-\mu|\geq k} k^2 f(y)dy$$
6. k^2 est une constante, nous pouvons la sortir de l'intégrale :
$$\sigma^2 \geq k^2 \int_{|y-\mu|\geq k} f(y)dy$$
7. Par définition, l'intégrale de la densité $f(y)$ sur la région $|y-\mu| \geq k$ n'est autre que la probabilité $P(|Y-\mu| \geq k)$.
$$\sigma^2 \geq k^2 \cdot P(|Y-\mu| \geq k)$$
8. En réarrangeant les termes (puisque $k > 0$, $k^2 > 0$), nous obtenons l'inégalité désirée :
$$P(|Y-\mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

Cette preuve est un cas particulier de l'Inégalité de Markov (appliquée à la variable aléatoire non-négative $X = (Y-\mu)^2$ et à la constante $a = k^2$).

Intuition : Comprendre l'Inégalité de Chebyshev

Cette formule peut être lue comme suit :
"La probabilité de s'écarter de la moyenne (μ) d'au moins k est bornée par la variance divisée par k^2."

- **Le rôle de la variance (σ^2) :** Si la variance est grande, la borne supérieure est élevée. L'inégalité nous dit "il est possible que la variable s'éloigne", ce qui est logique pour une grande dispersion. Si la variance est faible, la borne est basse, ce qui force la probabilité d'être loin à être faible.
- **Le rôle de l'écart (k) :** Le terme k^2 au dénominateur est crucial. Il signifie que la probabilité de s'écarter de la moyenne diminue *quadratiquement* avec la distance k. Être très loin est (relativement) très improbable.

Exemple : Une Borne Universelle

Exprimons l'inégalité en termes d'écarts-types (en posant $k = c \cdot \sigma$) :
$$P(|Y-\mu| \geq c\sigma) \leq \frac{\sigma^2}{(c\sigma)^2} = \frac{1}{c^2}$$

- **Pour $c = 2$:** $P(|Y-\mu| \geq 2\sigma) \leq \frac{1}{4} = 25\%$. Peu importe la distribution (symétrique, asymétrique, bizarre...), la probabilité d'être à 2 écarts-types ou plus de la moyenne est **au maximum** de 25%. (Pour une loi normale, cette probabilité est bien plus faible, $\approx 4.55\%$).
- **Pour $c = 3$:** $P(|Y-\mu| \geq 3\sigma) \leq \frac{1}{9} \approx 11.1\%$. La probabilité d'être à 3 écarts-types ou plus est au maximum de 11.1%. (Pour une loi normale, c'est $\approx 0.27\%$).

Chebyshev fournit une borne "garantie", bien que souvent non optimale. Elle est l'outil parfait pour la preuve qui suit.

9.2 La Loi Faible des Grands Nombres (LFGN / WLLN)

La loi faible stipule que la probabilité que notre moyenne d'échantillon s'écarte de la vraie moyenne de plus qu'une petite quantité ϵ tend vers zéro. C'est une **convergence en probabilité**.

Définition : Convergence en Probabilité

On dit qu'une suite de variables aléatoires Y_n converge en probabilité vers une constante c, noté $Y_n \xrightarrow{P} c$, si pour tout $\epsilon > 0$ (aussi petit soit-il) :

$$\lim_{n\to\infty} P(|Y_n - c| > \epsilon) = 0$$

Intuition : Comprendre la Convergence en Probabilité

La définition $P(|\bar{X}_n - \mu| > \epsilon) \to 0$ signifie :

- ϵ est votre **marge d'erreur** acceptable (ex : 0.01).
- $|\bar{X}_n - \mu|$ est l'erreur réelle de votre estimation.
- $P(\dots)$ est la probabilité que votre erreur **dépasse** votre marge.
- $\lim_{n\to\infty}(\dots) = 0$ signifie : "Si vous prenez un échantillon n suffisamment grand, la probabilité de faire une erreur plus grande que ϵ devient négligeable."

C'est une affirmation sur ce qui se passe pour un n fixe et très grand.

Théorème : Loi Faible des Grands Nombres (Khinchine)

Soit $X_1, X_2, \dots, X_n$ une suite de variables aléatoires **i.i.d.** (indépendantes et identiquement distribuées) avec une espérance finie $E[X_i] = \mu$. Soit $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ la moyenne d'échantillon. Alors, $\bar{X}_n$ converge en probabilité vers μ :

$$\bar{X}_n \xrightarrow{P} \mu$$

Preuve : Preuve (simplifiée) via l'Inégalité de Chebyshev

La loi faible de Khinchine ne nécessite qu'une moyenne finie. Cependant, si nous ajoutons la condition que la **variance σ^2 est aussi finie**, la preuve devient très simple.
Elle repose directement sur l'Inégalité de Chebyshev, que nous venons de voir. Nous l'appliquons à la variable aléatoire $Y = \bar{X}_n$.

1. Identifions les termes pour l'inégalité $P(|Y - E[Y]| \geq k) \leq \frac{\text{Var}(Y)}{k^2}$:
   - Notre variable est $Y = \bar{X}_n$.
   - Son espérance est $E[Y] = E[\bar{X}_n] = \mu$.
   - Sa variance est $\text{Var}(Y) = \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$.
   - Notre écart k est la marge d'erreur ϵ.
2. (Rappel du calcul de la variance de $\bar{X}_n$) : Puisque les X_i sont i.i.d., $\text{Var}(\bar{X}_n) = \text{Var}\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n^2}\sum \text{Var}(X_i) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$.
3. Appliquons l'inégalité de Chebyshev avec ces termes :

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\text{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2/n}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

4. Prenons maintenant la limite quand $n \to \infty$:

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

5. Puisque σ^2 et ϵ^2 sont des constantes finies, le terme de droite $\frac{\text{constante}}{n}$ tend vers 0.
6. Comme une probabilité ne peut pas être négative, nous avons :

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

C'est exactement la définition de la convergence en probabilité.

9.3 La Loi Forte des Grands Nombres (LFGN / SLLN)

La loi forte est une affirmation beaucoup plus puissante. Elle ne dit pas seulement qu'un "gros" écart est improbable pour un n "grand" ; elle dit que la probabilité que la suite $\bar{X}_n$ *ne converge pas* vers μ est nulle. C'est une **convergence presque sûre**.

Définition : Convergence Presque Sûre

On dit qu'une suite de variables aléatoires Y_n converge presque sûrement vers une constante c, noté $Y_n \xrightarrow{p.s.} c$, si :

$$P\left(\lim_{n\to\infty} Y_n = c\right) = 1$$

Théorème : Loi Forte des Grands Nombres (Kolmogorov)

Soit $X_1, X_2, \ldots, X_n$ une suite de variables aléatoires **i.i.d.** avec une espérance finie $E[X_i] = \mu$. Alors, $\bar{X}_n$ converge presque sûrement vers μ :

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

Remarque : Forte implique Faible

La convergence "presque sûre" (SLLN) est une condition plus stricte que la convergence "en probabilité" (WLLN). Si une suite converge presque sûrement, elle converge aussi en probabilité. L'inverse n'est pas toujours vrai.

9.4 Différence : Faible vs. Forte

Intuition : Faible vs. Forte : L'Analogie du Casino

Soit $\bar{X}_n$ votre gain moyen par partie après avoir joué n fois à la roulette. La vraie moyenne (l'avantage de la maison) est $\mu = -0.053$ (pour une roulette américaine).

- **Loi Faible (WLLN) :** "Si vous prévoyez de jouer $n = 1$ million de parties ce soir. La probabilité qu'à la fin de votre millionième partie, votre moyenne $\bar{X}_{1,000,000}$ soit loin de -0.053 (par exemple, que vous soyez gagnant, $\bar{X}_n > 0$) est infinitésimale."
- C'est une affirmation sur la distribution de $\bar{X}_n$ **à un point fixe** n **(très grand)**. Elle n'exclut pas la possibilité théorique (mais improbable) que si vous continuiez à jouer, votre moyenne $\bar{X}_n$ puisse à nouveau diverger follement avant de reconverger plus tard.
- **Loi Forte (SLLN) :** "Si vous jouez à la roulette *pour l'éternité*, en regardant la séquence de vos moyennes $\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots, \bar{X}_n, \ldots$."
- "La probabilité que cette **séquence entière** ne converge pas exactement vers $\mu = -0.053$ est de 0."
- C'est une affirmation sur la **trajectoire complète**. Elle dit que, avec une probabilité de 1, la trajectoire de $\bar{X}_n$ va s'approcher de μ et **ne plus s'en écarter** de manière significative.

En résumé :

- **Faible :** Pour n assez grand, un écart est **improbable**.
- **Forte :** La **trajectoire** converge vers μ (avec une probabilité de 1).

9.5 Application : La Méthode de Monte-Carlo

La Loi Forte des Grands Nombres est le moteur de l'une des techniques de calcul les plus puissantes : la simulation de Monte-Carlo. Elle nous permet d'estimer des quantités complexes (comme des intégrales) en utilisant le hasard.

Exemple : Estimer la valeur de π

Problème : Comment calculer π sans formule géométrique ?
Méthode (Statistique) :

1. Imaginez un carré de côté 1 (de $(0,0)$ à $(1,1)$). Son aire est $A_{\text{carré}} = 1$.

2. Imaginez un quart de cercle de rayon $r = 1$ inscrit dans ce carré. Son aire est $A_{\text{cercle}} = \frac{1}{4}\pi r^2 = \frac{\pi}{4}$.
3. Le *ratio* des aires est $\frac{A_{\text{cercle}}}{A_{\text{carré}}} = \frac{\pi/4}{1} = \frac{\pi}{4}$.

Simulation :

1. Nous allons "lancer des fléchettes" au hasard sur ce carré n fois.
2. Pour ce faire, nous générons n paires de nombres aléatoires (X_i, Y_i), où $X_i \sim U(0,1)$ et $Y_i \sim U(0,1)$.
3. Pour chaque point i, nous vérifions s'il a atterri **dans le cercle**. La condition est $X_i^2 + Y_i^2 \leq 1$.
4. Nous définissons une nouvelle variable aléatoire Z_i (de Bernoulli) :

$$Z_i = \begin{cases} 1 & \text{si } X_i^2 + Y_i^2 \leq 1 \quad \text{(le point est dans le cercle)} \\ 0 & \text{sinon} \end{cases}$$

Application de la LLN :

- Quelle est la "vraie moyenne" μ de cette variable Z_i ?
- $\mu = E[Z_i] = 1 \cdot P(Z_i = 1) + 0 \cdot P(Z_i = 0) = P(Z_i = 1)$.
- $P(Z_i = 1)$ est la probabilité qu'un point aléatoire tombe dans le cercle. Puisque les points sont uniformes, cette probabilité est simplement le ratio des aires !
- Donc, la vraie moyenne (inconnue) est $\mu = \frac{A_{\text{cercle}}}{A_{\text{carré}}} = \frac{\pi}{4}$.
- Comment estimer μ ? Nous utilisons la moyenne d'échantillon $\bar{Z}_n$:

$$\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i = \frac{\text{Nombre de points dans le cercle}}{n}$$

- Par la **Loi Forte des Grands Nombres**, nous avons la garantie que :

$$\bar{Z}_n \xrightarrow{p.s.} \mu = \frac{\pi}{4}$$

Conclusion : Pour estimer π, il suffit de calculer $\bar{Z}_n$ (une simple proportion) et de la multiplier par 4.

$$\pi \approx 4 \cdot \bar{Z}_n$$

Plus notre nombre de simulations n est grand, plus la SLLN nous garantit que notre estimation sera proche de la vraie valeur de π.

10 Le Théorème Central Limite (TCL)

10.1 Introduction : L'omniprésence de la loi normale

Dans la section précédente, la Loi des Grands Nombres (LLN) nous a donné une garantie fondamentale : la moyenne d'échantillon $\bar{X}_n$ converge vers la vraie moyenne μ lorsque n devient grand.

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

La LLN nous dit **où** la moyenne d'échantillon converge (vers la constante μ), mais elle ne nous dit rien sur la *forme* de la distribution de $\bar{X}_n$ autour de μ pour un n grand, mais fini.

Le **Théorème Central Limite (TCL)** comble cette lacune. Il décrit la *manière* dont $\bar{X}_n$ converge, en nous donnant la forme de sa distribution. C'est sans doute le théorème le plus important des statistiques.

Intuition : L'Idée Fondamentale

Intuitivement, ce résultat affirme qu'une **somme** d'un grand nombre de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) tend, le plus souvent, à suivre une **loi normale** (aussi appelée loi de Laplace-Gauss ou "courbe en cloche").
Ce théorème et ses généralisations offrent une explication à l'omniprésence de la loi normale dans la nature. De nombreux phénomènes (la taille d'un individu, l'erreur de mesure d'un instrument, le bruit de fond d'un signal) sont le résultat de l'addition d'un très grand nombre de petites perturbations aléatoires. Le TCL nous dit que le résultat de cette somme sera, inévitablement, distribué selon une loi normale.

10.2 L'illustration : la somme des "Pile ou Face"

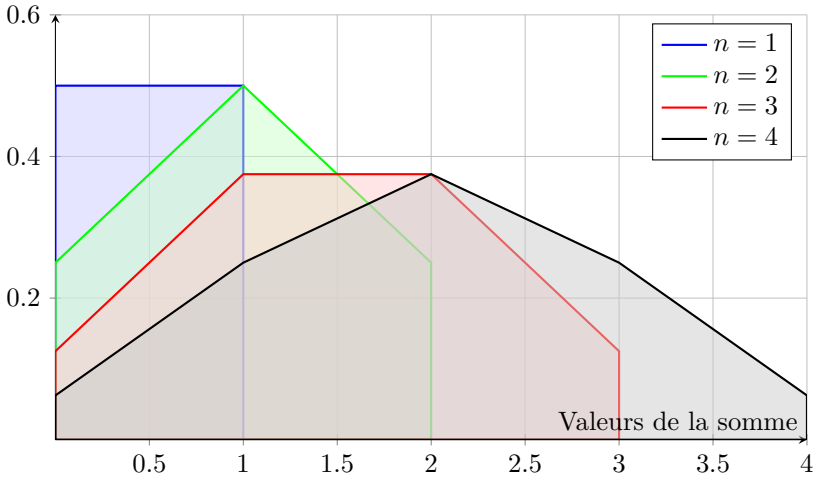
Prenons l'exemple le plus simple pour illustrer ce phénomène : le jeu de "pile ou face".

Exemple : Distribution de la Somme de n Lancers

Soit X_i le résultat du i-ème lancer, avec $X_i = 1$ pour "Face" (probabilité 0,5) et $X_i = 0$ pour "Pile" (probabilité 0,5). La distribution d'origine (pour $n = 1$) n'est pas du tout une courbe en cloche : c'est une distribution discrète avec deux bâtons de même hauteur.
Considérons la **somme** $S_n = X_1 + X_2 + \cdots + X_n$, qui représente le nombre total de "Face" obtenus en n lancers.

- **Pour $n = 1$:** La distribution de S_1 est :
  - Valeurs de la somme : $\{0, 1\}$
  - Fréquences : $\{0.5, 0.5\}$
- **Pour $n = 2$:** Les sommes possibles sont $\{0, 1, 2\}$. La distribution de S_2 est :
  - Valeurs de la somme : $\{0, 1, 2\}$
  - Fréquences : $\{0.25, 0.5, 0.25\}$ (elle forme un triangle).
- **Pour $n = 3$:** Les sommes possibles sont $\{0, 1, 2, 3\}$. La distribution de S_3 est :
  - Valeurs de la somme : $\{0, 1, 2, 3\}$
  - Fréquences : $\{0.125, 0.375, 0.375, 0.125\}$

Fonction de fréquence pour des tirages à pile ou face



Graphiquement, on constate que plus le nombre de tirages n augmente (par exemple, jusqu'à $n = 12$), plus la courbe de fréquence (qui reste discrète) se rapproche d'une courbe en cloche symétrique, caractéristique de la loi normale.

10.3 Distribution de la population vs. Distribution d'échantillonnage

Le point le plus remarquable du TCL est qu'il fonctionne *quelle que soit* la distribution de départ.

Intuition : Population vs. Échantillonnage

Imaginez deux univers de distributions :

- **1. La Distribution de la Population (X_i) :** C'est la loi de nos variables X_i individuelles. Elle peut avoir **n'importe quelle forme** (par exemple, une distribution bimodale, asymétrique, ou uniforme). Cette distribution a une "vraie" moyenne μ et un "vrai" écart-type σ.
- **2. La Distribution d'Échantillonnage ($\bar{X}_n$) :** C'est la distribution de la *moyenne* $\bar{X}_n = (X_1 + \cdots + X_n)/n$, calculée sur des échantillons de taille n. C'est la distribution de "toutes les moyennes d'échantillon possibles".

Le TCL énonce la relation magique entre les deux :
Quelle que soit la forme de la distribution de la population, plus la taille de l'échantillon n croît, plus la distribution d'échantillonnage de la moyenne $\bar{X}_n$ est proche d'une loi normale (gaussienne).
De plus, les paramètres de cette loi normale sont :

- **Moyenne :** La distribution de $\bar{X}_n$ est centrée sur la même moyenne μ que la population.
- **Écart-type :** La distribution de $\bar{X}_n$ est beaucoup plus resserrée. Son écart-type (appelé "erreur standard") est $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

Cette dispersion $\sigma/\sqrt{n}$ qui tend vers 0 est la manifestation de la Loi des Grands Nombres. Le TCL précise que la *forme* de cette convergence est gaussienne.

10.4 Énoncé formel du Théorème Central Limite

Pour énoncer le théorème formellement, nous devons d'abord définir les propriétés de la somme S_n et de la moyenne $\bar{X}_n$.

Soit $X_1, \ldots, X_n$ des variables aléatoires i.i.d. avec $E[X_i] = \mu$ et $\text{Var}(X_i) = \sigma^2$.

- **La Somme $S_n = \sum X_i$:**
  - Espérance : $E[S_n] = E[\sum X_i] = \sum E[X_i] = n\mu$
  - Variance : $\text{Var}(S_n) = \text{Var}(\sum X_i) = \sum \text{Var}(X_i) = n\sigma^2$
  - Écart-type : $\sigma_{S_n} = \sqrt{n\sigma^2} = \sigma\sqrt{n}$

· **La Moyenne** $\bar{X}_n = S_n/n$:
  — Espérance : $E[\bar{X}_n] = E[S_n/n] = \frac{1}{n}E[S_n] = \frac{1}{n}(n\mu) = \mu$
  — Variance : $\text{Var}(\bar{X}_n) = \text{Var}(S_n/n) = \frac{1}{n^2}\text{Var}(S_n) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$
  — Écart-type : $\sigma_{\bar{X}_n} = \sqrt{\sigma^2/n} = \frac{\sigma}{\sqrt{n}}$

Nous voyons que la distribution de S_n s'étale (variance $\to \infty$) tandis que celle de $\bar{X}_n$ se contracte (variance $\to 0$). Pour étudier la *forme* de la convergence, nous créons une variable "stable" en la centrant (soustrayant la moyenne) et en la réduisant (divisant par l'écart-type). C'est la variable Z_n.

Théorème : Théorème Central Limite (Lindeberg-Lévy)

Soit $X_1, X_2, \ldots, X_n$ une suite de variables aléatoires **i.i.d.** (indépendantes et identiquement distribuées) suivant la même loi D. Supposons que l'**espérance** μ et l'**écart-type** σ de cette loi D existent, sont finis, et $\sigma \neq 0$.
Considérons la variable aléatoire standardisée Z_n :

$$Z_n = \frac{S_n - E[S_n]}{\sigma_{S_n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Cette variable est équivalente à la moyenne standardisée :

$$Z_n = \frac{\bar{X}_n - E[\bar{X}_n]}{\sigma_{\bar{X}_n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

(Pour tout n, Z_n est une variable centrée-réduite : $E[Z_n] = 0$ et $\text{Var}(Z_n) = 1$).
Alors, la suite de variables aléatoires $Z_1, Z_2, \ldots, Z_n, \ldots$ **converge en loi** vers une variable aléatoire Z qui suit la **loi normale centrée réduite** $N(0,1)$, lorsque n tend vers l'infini.
Cela signifie que si Φ est la fonction de répartition de la loi $N(0,1)$, alors pour tout réel z :

$$\lim_{n\to\infty} P(Z_n \leq z) = \lim_{n\to\infty} P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) = \Phi(z)$$

11 Appendice A : Séries de Taylor et Maclaurin

Définition : Séries de Taylor et Maclaurin

Si une fonction f est indéfiniment dérivable au voisinage d'un point a, sa **série de Taylor** centrée en a est définie par :

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k$$

où $f^{(k)}(a)$ est la k-ième dérivée de f évaluée en a.

Dans le cas particulier où $\mathbf{a = 0}$, la série est appelée une **série de Maclaurin**. C'est la forme la plus courante, car elle approxime les fonctions autour de l'origine.

11.1 Construction pas à pas d'une série de Taylor

Intuition : La logique de la correspondance des dérivées

L'objectif fondamental d'une série de Taylor est de construire un polynôme, $P(x)$, qui soit une "copie conforme" d'une fonction $f(x)$ autour d'un point a. Pour ce faire, on force le polynôme à avoir exactement les mêmes propriétés locales que la fonction : même valeur, même pente, même courbure, etc. Cela se traduit mathématiquement par une exigence : **la n-ième dérivée du polynôme en a doit être égale à la n-ième dérivée de la fonction en a**, et ce pour tous les ordres n.
Prenons l'exemple de $f(x) = e^x$ et construisons sa série de Maclaurin (centrée en $a = 0$), où $f^{(k)}(0) = 1$ pour tout k.

1. **Ordre 0 : Faire correspondre la valeur**
   Objectif : Le polynôme $P_0(x)$ doit avoir la même valeur que $f(x)$ en $x = 0$. On veut $P_0(0) = f(0)$.
   Solution : On choisit le polynôme le plus simple, une constante : $P_0(x) = f(0)$. Pour e^x, $f(0) = 1$, donc $\mathbf{P_0(x) = 1}$.
   Vérification : $P_0(0) = 1$. L'objectif est atteint.
2. **Ordre 1 : Faire correspondre la première dérivée**
   Objectif : On veut un nouveau polynôme $P_1(x)$ qui préserve la correspondance précédente ($P_1(0) = f(0)$) ET qui a la même pente, c'est-à-dire $P_1'(0) = f'(0)$.
   Solution : On ajoute un terme en x à notre polynôme précédent : $P_1(x) = P_0(x) + c_1x = 1 + c_1x$.
   Vérification :
   - $P_1(0) = 1 + c_1(0) = 1$. La valeur correspond toujours, car le nouveau terme s'annule en 0.
   - On dérive : $P_1'(x) = c_1$. Pour que les pentes correspondent en 0, il faut $P_1'(0) = c_1 = f'(0)$. Comme $f'(0) = 1$, on doit choisir $\mathbf{c_1 = 1}$.

   Notre polynôme est maintenant $\mathbf{P_1(x) = 1 + x}$.
3. **Ordre 2 : Faire correspondre la deuxième dérivée**
   Objectif : On veut $P_2(x)$ tel que $P_2(0) = f(0)$, $P_2'(0) = f'(0)$ ET $P_2''(0) = f''(0)$.
   Solution : On ajoute un terme en x^2 : $P_2(x) = P_1(x) + c_2x^2 = 1 + x + c_2x^2$.
   Vérification :
   - Les dérivées d'ordre 0 et 1 en $x = 0$ ne sont pas affectées, car la dérivée de c_2x^2 (soit $2c_2x$) et le terme lui-même s'annulent en 0. Les objectifs précédents sont préservés.
   - On dérive deux fois : $P_2'(x) = 1 + 2c_2x$ et $P_2''(x) = 2c_2$.
   - Pour que les courbures correspondent, il faut $P_2''(0) = 2c_2 = f''(0)$. Comme $f''(0) = 1$, on doit choisir $\mathbf{c_2 = 1/2}$.

   Notre polynôme est $\mathbf{P_2(x) = 1 + x + \frac{1}{2}x^2}$.
4. **Le schéma général : L'importance de la factorielle**
   Pour faire correspondre la k-ième dérivée, on ajoute un terme c_kx^k.

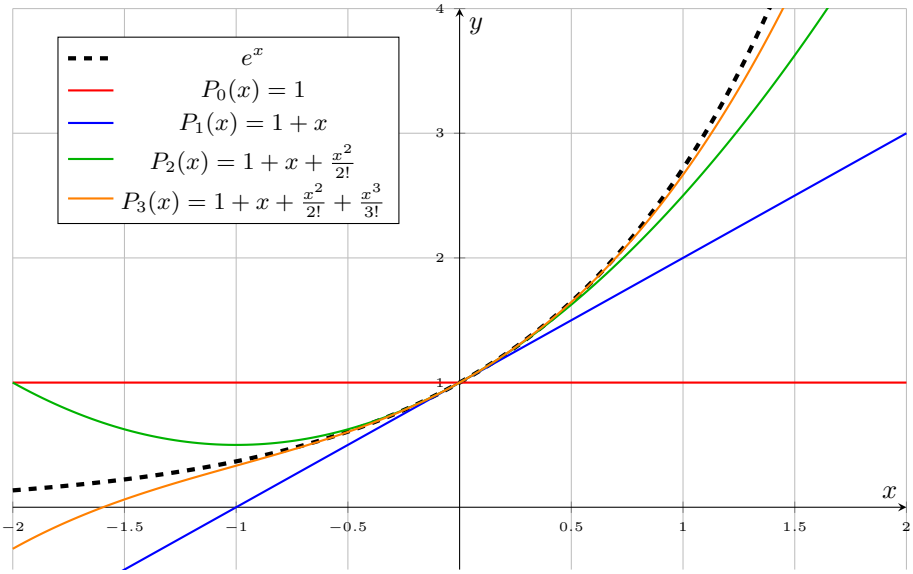
Quand on dérive $c_k x^k$ exactement k fois, on obtient $c_k \times k!$.
Toutes les dérivées d'ordre inférieur s'annulent en $x = 0$. On doit donc avoir :

$$P_k^{(k)}(0) = c_k \cdot k! = f^{(k)}(0)$$

Cela nous donne la règle pour trouver chaque coefficient :

$$c_k = \frac{f^{(k)}(0)}{k!}$$

C'est précisément le coefficient qui apparaît dans la formule de Taylor, et il est choisi pour cette unique raison : forcer la k-ième dérivée du polynôme à correspondre parfaitement à celle de la fonction au point de développement.



Visualisation de la construction progressive de la série de Maclaurin pour e^x.

11.2 Intuition de la série de Taylor en un point quelconque a

Intuition : Construire une approximation loin de l'origine

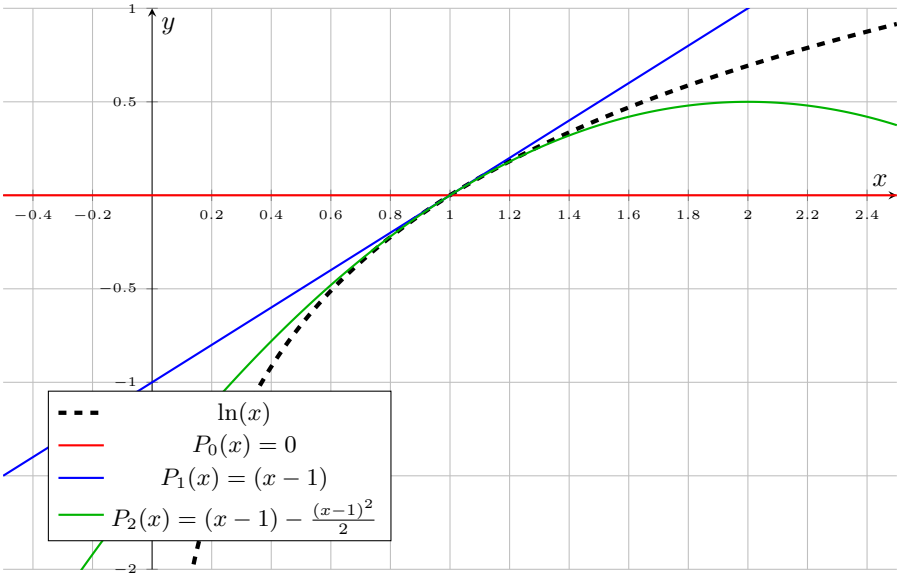
La série de Maclaurin est puissante, mais elle nous contraint à approximer une fonction uniquement autour de $x = 0$. Que faire si l'on s'intéresse au comportement d'une fonction ailleurs, par exemple $f(x) = \ln(x)$ autour de $x = 1$ (puisque $\ln(0)$ n'est pas défini) ? C'est là qu'intervient la série de Taylor générale.
L'objectif reste le même : construire un polynôme $P(x)$ qui est une "copie conforme" de $f(x)$ au point a. Pour cela, on force les dérivées du polynôme à correspondre à celles de la fonction en ce point a. La seule différence est que notre "variable" de base n'est plus x, mais l'écart par rapport au centre, c'est-à-dire $(x - a)$.
Prenons l'exemple de $f(x) = \ln(x)$ et construisons sa série centrée en $\mathbf{a = 1}$.

1. **Ordre 0 : Faire correspondre la valeur**
   Objectif : $P_0(a) = f(a)$.
   Solution : On calcule $f(1) = \ln(1) = 0$. Le polynôme est la constante $\mathbf{P_0(x) = 0}$.
2. **Ordre 1 : Faire correspondre la pente**
   Objectif : $P_1(a) = f(a)$ et $P_1'(a) = f'(a)$.
   Solution : On ajoute un terme proportionnel à l'écart $(x-a)$: $P_1(x) = f(a) + c_1(x-a)$.
   Vérification :
   - $P_1(1) = 0 + c_1(1 - 1) = 0$. La valeur correspond.
   - On dérive : $P_1'(x) = c_1$. On veut $P_1'(1) = c_1 = f'(1)$.
   - La dérivée de $f(x) = \ln(x)$ est $f'(x) = 1/x$, donc $f'(1) = 1$. On doit choisir $\mathbf{c_1 = 1}$.

   Notre polynôme est $\mathbf{P_1(x) = (x - 1)}$. C'est la tangente à $\ln(x)$ en $x = 1$.

3. **Ordre 2 : Faire correspondre la courbure**
   Objectif : Les dérivées jusqu'à l'ordre 2 doivent correspondre en $a = 1$.
   Solution : On ajoute un terme en $(x-a)^2$: $P_2(x) = (x-1) + c_2(x-1)^2$.
   Vérification :
   - Les correspondances d'ordre 0 et 1 sont préservées.
   - On dérive deux fois : $P_2'(x) = 1 + 2c_2(x-1)$ et $P_2''(x) = 2c_2$.
   - On veut $P_2''(1) = 2c_2 = f''(1)$.
   - La dérivée seconde de $f(x)$ est $f''(x) = -1/x^2$, donc $f''(1) = -1$. On choisit $\mathbf{c_2 = -1/2}$.

   Notre polynôme est $\mathbf{P_2(x) = (x-1) - \frac{1}{2}(x-1)^2}$.
4. **Le schéma général**
   Le coefficient c_k du terme $(x-a)^k$ est choisi pour faire correspondre la k-ième dérivée. La dérivation de $c_k(x-a)^k$ k fois donne $c_k \cdot k!$. On impose donc $c_k \cdot k! = f^{(k)}(a)$, ce qui mène directement à la formule générale $c_k = \frac{f^{(k)}(a)}{k!}$.



Approximation de $\ln(x)$ *autour de* $a = 1$. *Le polynôme "colle" à la fonction près de* $x = 1$.

11.3 La Fonction Exponentielle (e^x)

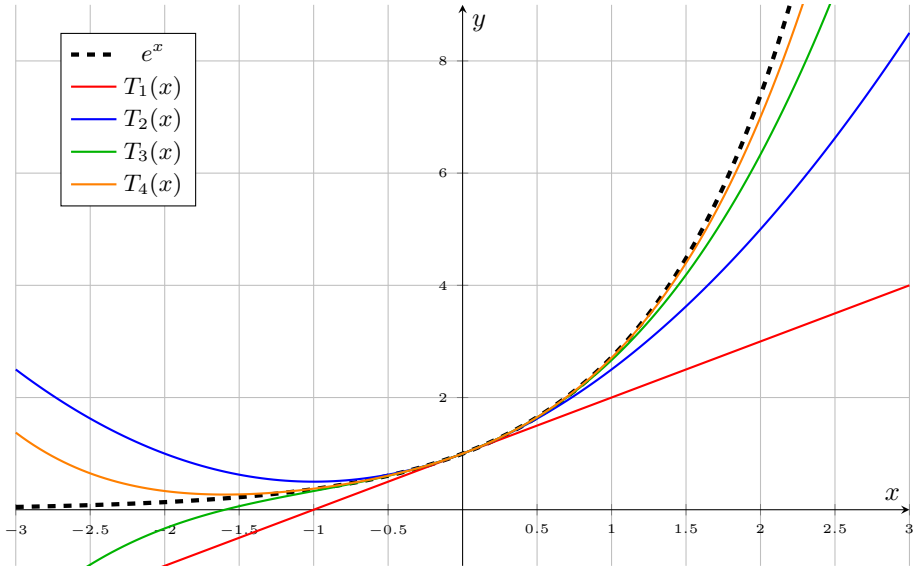
Théorème : Série de Maclaurin pour e^x

Pour tout nombre réel x, la fonction exponentielle peut s'écrire :

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

Intuition : Visualiser la Croissance Exponentielle

La fonction exponentielle est unique car elle est sa propre dérivée. Cela signifie que toutes ses informations locales (valeur, pente, courbure) en $a = 0$ sont égales à **1**. La série pour e^x est donc le polynôme le plus « pur », où chaque terme x^k est simplement normalisé par $k!$. Le graphique ci-dessous montre comment les polynômes de Taylor convergent rapidement vers la véritable courbe exponentielle, illustrant sa croissance puissante.



Approximation de e^x par ses polynômes de Maclaurin.

Preuve

Soit $f(x) = e^x$. Pour tout entier $k \geq 0$, la k-ième dérivée est $f^{(k)}(x) = e^x$. En évaluant en $a = 0$, on obtient $f^{(k)}(0) = e^0 = 1$ pour tout k. En appliquant la formule de Maclaurin :

$$e^x = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k = \sum_{k=0}^{\infty} \frac{1}{k!} x^k = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots$$

11.4 La Fonction Sinus $(\sin(x))$

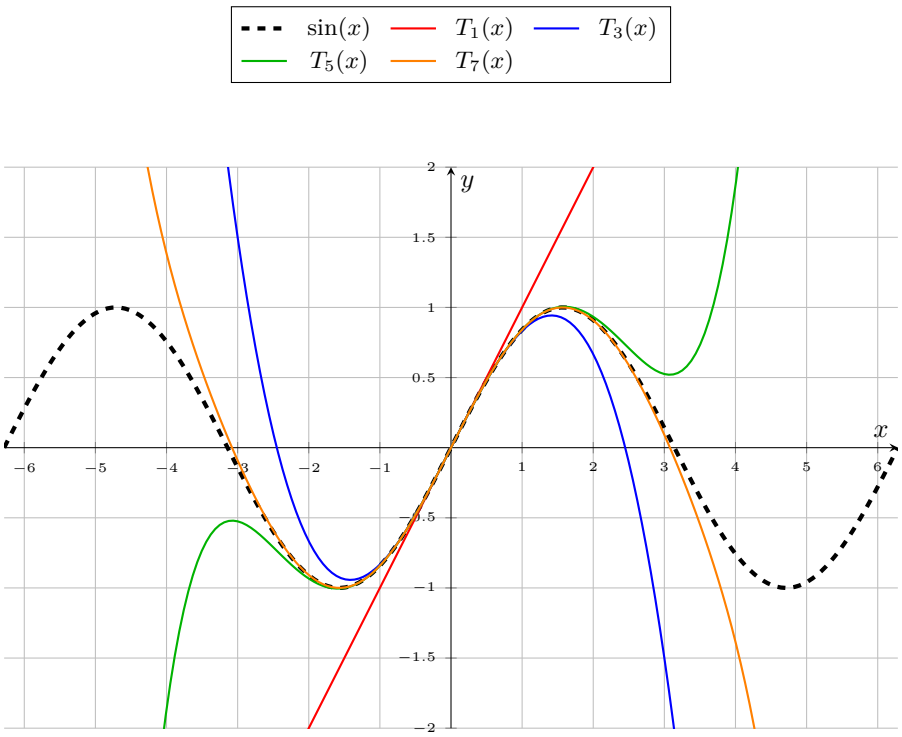
Théorème : Série de Maclaurin pour $\sin(x)$

Pour tout nombre réel x :

$$\sin(x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

Intuition : Visualiser l'Oscillation du Sinus

La série du sinus reflète ses propriétés fondamentales. En tant que fonction **impaire** ($\sin(-x) = -\sin(x)$), son développement ne contient que des puissances **impaires** de x. Les signes alternés capturent sa nature oscillatoire. Le graphique ci-dessous montre comment l'ajout de termes permet au polynôme d'« épouser » la courbe du sinus sur un plus grand nombre de périodes.



Approximation de $\sin(x)$ *par ses polynômes de Maclaurin.*

Preuve

Soit $f(x) = \sin(x)$. Les dérivées en $a = 0$ suivent un cycle $(0, 1, 0, -1, \dots)$. Seuls les termes d'ordre impair $(2k+1)$ sont non nuls, avec des valeurs de $(-1)^k$, ce qui donne la formule.

11.5 La Fonction Cosinus ($\cos(x)$)

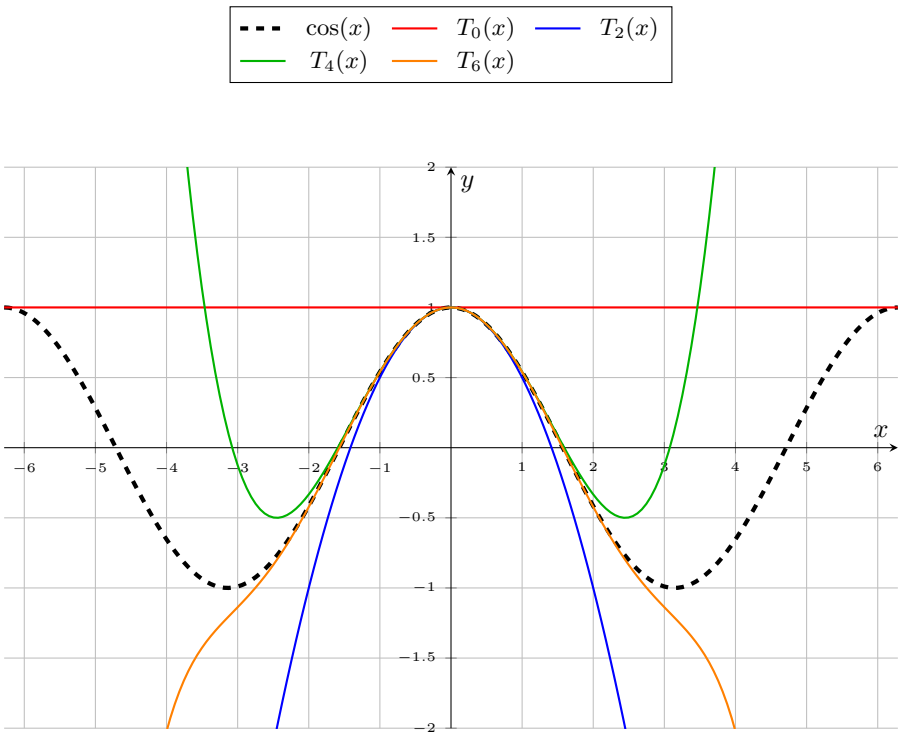
Théorème : Série de Maclaurin pour $\cos(x)$

Pour tout nombre réel x :

$$\cos(x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots$$

Intuition : Visualiser la Symétrie du Cosinus

En tant que fonction **paire** ($\cos(-x) = \cos(x)$), la série du cosinus ne contient, de manière appropriée, que des puissances **paires** de x. Elle commence à 1 (son maximum) puis oscille, un comportement capturé par les signes alternés.



Approximation de $\cos(x)$ *par ses polynômes de Maclaurin.*

Preuve

Soit $g(x) = \cos(x)$. Les dérivées en $a = 0$ suivent un cycle $(1, 0, -1, 0, \dots)$. Seuls les termes d'ordre pair $(2k)$ sont non nuls, avec des valeurs de $(-1)^k$, ce qui donne la formule.

11.6 Le Logarithme Népérien $(\ln(1+x))$

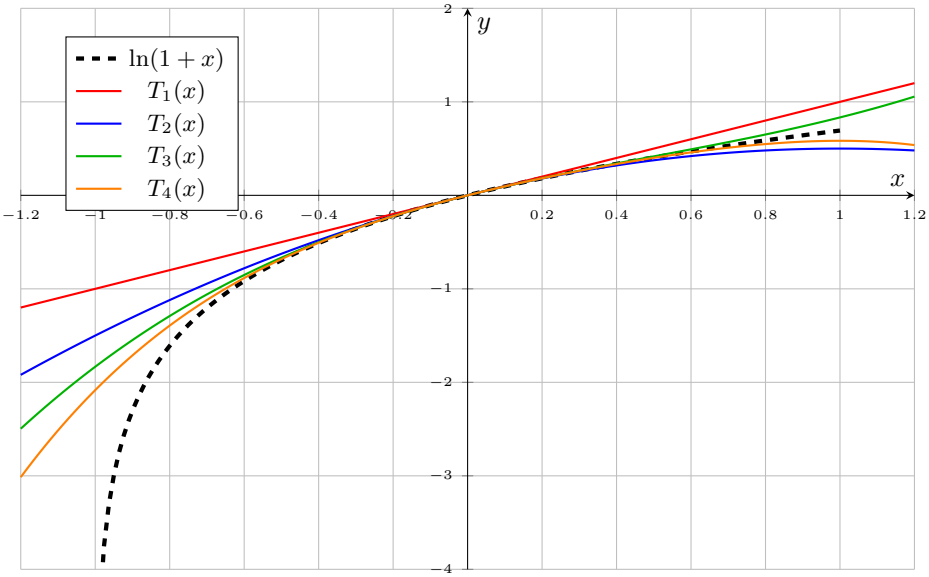
Théorème : Série de Maclaurin pour $\ln(1+x)$

Pour $|x| < 1$:

$$\ln(1+x) = \sum_{k=1}^{\infty} (-1)^{k-1} \frac{x^k}{k} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots$$

Intuition : Visualiser l'Approximation Logarithmique

Cette série est essentielle pour approximer les logarithmes près de 1. Contrairement aux fonctions précédentes, elle ne converge que pour $|x| < 1$. Le graphique montre que l'approximation est excellente près de $x = 0$ mais diverge rapidement lorsque x s'approche de la frontière de convergence à $x = 1$.



Approximation de $\ln(1+x)$ *par ses polynômes de Maclaurin.*

Preuve

Soit $f(x) = \ln(1+x)$. Pour $k \geq 1$, la k-ième dérivée en $a = 0$ est $f^{(k)}(0) = (-1)^{k-1}(k-1)!$. En substituant cela dans la formule de Maclaurin, le $(k-1)!$ au numérateur annule partiellement le $k!$ au dénominateur, laissant un k en bas.

11.7 La Série Géométrique ($\frac{1}{1-x}$)

Théorème : Série de Maclaurin pour $\frac{1}{1-x}$

Pour $|x| < 1$:

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k = 1 + x + x^2 + x^3 + \cdots$$

Intuition : Le Fondement de Nombreuses Séries

Cette série, connue sous le nom de série géométrique, est l'un des développements en série de puissances les plus fondamentaux. Elle converge uniquement lorsque la valeur absolue de x est inférieure à 1. Chaque coefficient est simplement 1, ce qui en fait la série de Maclaurin la plus simple. De nombreuses autres séries, comme celle de $\ln(1+x)$ ou de $\arctan(x)$, peuvent être dérivées de celle-ci par intégration ou substitution.

Preuve

Soit $f(x) = (1-x)^{-1}$. Les dérivées successives sont $f'(x) = 1(1-x)^{-2}$, $f''(x) = 2(1-x)^{-3}$, $f'''(x) = 6(1-x)^{-4}$, et ainsi de suite. La formule générale pour la k-ième dérivée est $f^{(k)}(x) = k!(1-x)^{-(k+1)}$. En évaluant en $a = 0$, on obtient $f^{(k)}(0) = k!$. En substituant dans la formule de Maclaurin :

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k = \sum_{k=0}^{\infty} \frac{k!}{k!} x^k = \sum_{k=0}^{\infty} x^k$$