Probabilités et Statistique (MATH233)

Victor Panaretos

Institut de Mathématiques – EPFL

victor.panaretos@epfl.ch

- Cours mardi 13h15–15h00
- Exercices lundi 09h15–10h00
- Livres (open access!) de référence principaux :

  Dalang & Conus, *Introduction à la théorie des probabilités*, PPUR

  Panaretos, *Statistique pour Mathématiciens*, PPUR
- Page web : moodle
- Test midterm facultatif (bonus), formulaire manuscrit A4 recto autorisé

  18 novembre, 13h15-15h00
- Examen final écrit (formulaire manuscrit A4 recto-verso autorisé)
- La note finale G sera calculée comme suit :
  - $F = 0.6 \times E + 0.4 \times \max\{E, T\}$
  - $E =$ examen, $T =$ test
  - on arrondit F pour obtenir G

Introduction

En bref :

la quantification rigoureuse de l'incertitude

The scientist has a lot of experience with ignorance and doubt and uncertainty, and this experience is of very great importance[...] in order to progress we must recognize our ignorance and leave room for doubt. Scientific knowledge is a body of statements of varying degrees of certainty — some most unsure, some nearly sure, but none absolutely certain.

Richard Feynman

We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression.

Ronald A. Fisher

The object of rigor is to sanction and legitimize the the conquests of intuition, and there was never any other object for it.

Jacques Hadamard

L'incertitude peut avoir plusieurs sources :

1. Erreur de mesure.
2. Chaos.
3. Stochasticité intrinsèque.
4. Données échantillonnées (on observe le *particulier* mais pas le *général*).
5. Limites fondamentales à la précision.

⋮

On va encapsuler mathématiquement l'incertitude, peu importe la source !

Probabilités :

1. Le phénomène d'intérêt est conceptualisé comme un modèle probabiliste
2. On utilise le modèle pour apprendre la probabilité des résultats possibles.
3. La probabilité est le langage de la modélisation scientifique

Statistique :

1. Le phénomène d'intérêt est conceptualisé comme un modèle probabiliste
2. On utilise des données pour apprendre quelque chose sur le modèle.
3. La statistique est le langage de l'expérimentation scientifique

- Quels sont les comportements typiques et atypiques de ce modèle ?
  (par ex. concentration de la mesure, moments, événements rares)
- Comment se comportent les fluctuations autour du resultat "typique"
  (par ex. variance, phénomènes limites centrés, grandes déviations)
- Quel comportement ce modèle prédit-il à grande échelle ou à la limite ?
  (par ex. lois des grands nombres, transitions de phase, convergence)
- Quelles propriétés structurelles ce modèle présente-t-il ?
  (par ex. ergodicité, symétrie, invariance, indépendance)
- A quel point est le modèle stable face à des perturbations de ses ingrédients ?
  (par ex. classes d'universalité, stabilité sous dynamique)
- Comment des composantes aléatoires interagissent-elles dans un système complexe ?
  (par ex. systèmes de particules en interaction, champs aléatoires, percolation)

- Étant donnée plusierus modèles plausible, peut-on déterminer lequel a généré les données ?
  (estimation, choix de modèle)
- Les données sont-elles plus cohérentes avec un modèle qu'un autre ?
  (tests d'hypothèse, discrimination)
- Quel ensemble de modèles est cohérent avec un jeu de données donné ?
  (régions de confiance, postérieurs bayésiens)
- Comment répondre au mieux à ces questions — et une réponse optimale existe-t-elle ?
  (optimalité/efficacité, admissibilité, taux de convergence)
- Comment peut-on quantifier l'incertitude dans nos conclusions ?
  (lois d'echantillonage, écart type, distributions postérieures)
- Quels sont les risques d'erreur — et comment peut-on les contrôler ?
  (erreurs de type I/II, mauvaise spécification du modèle, robustesse)

Example (Un Probabiliste et un Statisticien lancent une pièce)

Soient $Y_1, ..., Y_n$ les résultats du lancer de pièce n fois, $Y_i = \begin{cases} 0 & \text{si face,} \\ 1 & \text{si pile.} \end{cases}$

Un modèle plausible est que chaque résultat Y_i a la même probabilité $\theta \in [0, 1]$ de'être 0 ou 1, et qu'aucun résultat n'influence les autres. **Questions de proba :**

- Quelle est la probabilité de la séquence $(0, 0, 0, 1, 0, 1, 1, 1, 1, 1)$ en fonction de θ ?
- L'ordre des 0 et des 1 joue-t-il un rôle, ou observe-t-on une forme d'invariance ?
- Quelle est la probabilité de l'apparition d'une suite de zéros de longueur k ($\leq n$) ?
- Si l'on continue à lancer indéfiniment, combien de suites de longueur k apparaîtront ? Et combien de temps avant la première ?
- Combien de fois une séquence spécifique (par exemple 0,1,0,1,0,1,0,1,0,1,0) apparaîtra-t-elle ? Comment cela dépend-il de la valeur de θ ?
- Que dire de la somme des observations ? Quel est son comportement ? Comment évolue-t-elle à grande échelle ($n \to \infty$) ?
- Si l'on trace la courbe $f_n(t) = \alpha_n \sum_{j=1}^{\lfloor nt \rfloor} (Y_j - \theta)$, à quoi ressemblera-t-elle lorsque $n \to \infty$? Quel rééchelonnement α_n faut-il choisir ?
- Et si l'on autorise les essais successifs à dépendre les uns des autres ?

Example (Un Probabiliste et un Statisticien lancent une pièce (suite))

Étant donné un certain résultat pour n lancers — par exemple $(0, 0, 0, 1, 0, 1, 1, 1, 1, 1)$ avec $n = 10$— quelques **questions statistiques** sont :

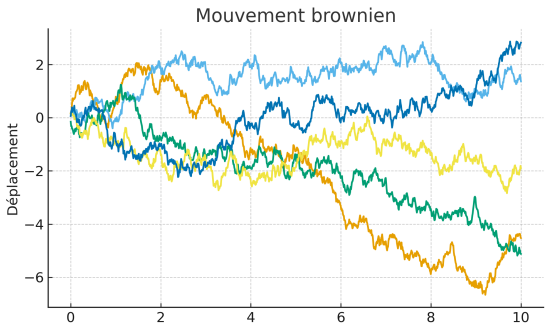
- La pièce est-elle équilibrée ?
- Quelle est une bonne estimation de la valeur de θ à partir des observations ?
- Quel intervalle de valeurs de θ est plausible selon les observations ?
- Quelle est l'erreur commise en tirant ces conclusions des observations ?
- Comment nos réponses changeraient-elles si les observations étaient légèrement modifiées ?
- Existe-t-il une solution "optimale" à ces problèmes ?
- Quelle est la sensibilité de nos réponses à des écarts au modèle ?
- Comment nos réponses évoluent-elles lorsque le nombre de lancers $\longrightarrow \infty$?
- Combien de lancers faudrait-il pour obtenir des "réponses précises" ?
- Est-ce qu'on peut juger si il y a une dependence séquentielle entre les résultats?
- Si chaque résultat a sa propre probabilité de succès θ_i, vaut-il la peine d'utiliser l'ensemble des résultats pour estimer les θ_i, ou faut-il traiter chaque résultat séparément ?

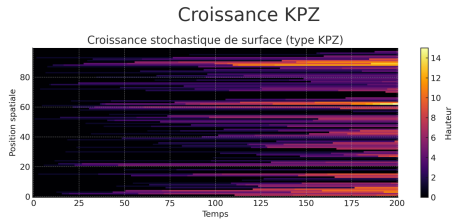
Cela paraît simpliste ? Peut déjà produire des réponses surprenantes.
(par ex.: singe tapant Shakespeare, phénomène de Stein)

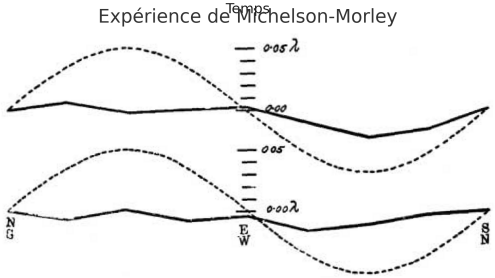
Et certaines de ces questions peuvent avoir une grande portée

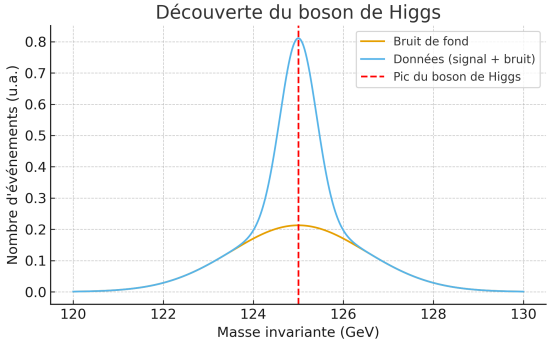
- Einstein/Wiener et le mouvement brownien (hypothèse atomique)
- Ising/. . . /Duminil-Copin et les systèmes de particules en interaction (ferromagnétisme)
- KPZ/ . . . /Hairer et la croissance aléatoire de surfaces (croissance cristalline)
- Michelson/Morley et l'éther luminifère (vitesse de la lumière)
- Higgs/. . . et son boson (modèle standard)
- . . . /Smoot/Mather et le rayonnement cosmique de fond (big bang)

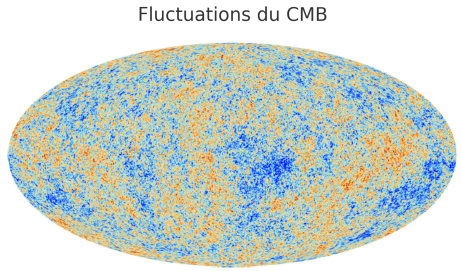
Probabilité et statistiques en physique : six jalons











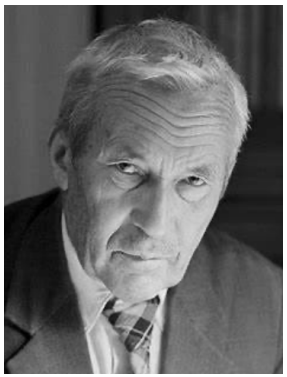
Pensons à la géométrie et contrastons l'approche descriptive vs prescriptive:

- Mathématiques babyloniennes et égyptiennes
- Système axiomatique formel d'Euclide

Il en va de même pour la probabilité — des approches descriptives ont été tentées:

- Laplace et le rapport cas favorables / cas totaux
- von Mises et la fréquence relative à long terme
- Buffon et l'approche géométrique/symétrique

Dans les années 1930, le moment était propice — et les outils (la théorie de la mesure) étaient disponibles :



The theory of probability as mathematical discipline can and should be developed from axioms in exactly the same way as Geometry and Algebra

Andrei Kolmogorov
Foundations of the Theory of Probability (1933)

L'un des grands moments de l'histoire des maths — et même de la science.

Dans la fondation axiomatique de la géométrie, nous avons :

- des points
- des droites et des angles droits
- des cercles

Quelles sont les "choses" de la probabilité ?

- Un ensemble Ω de tous les résultats possibles ω
- Une collection $\mathcal{F}$ d'observables possibles (événements) $A \subseteq \Omega$.
- Une application $\mathbb{P}$ qui associe à chaque événement une valeur de probabilité.

Quels sont quelques desiderata ? Nous voulons :

- que Ω soit sans restriction (dénombrable, non dénombrable, de dimension finie/infini)
- pouvoir formuler des propositions en combinant convenablement les observables
- que $\mathbb{P}$ respecte l'intuition de comptage/volume dans les espaces dénombrables/non dénombrables.

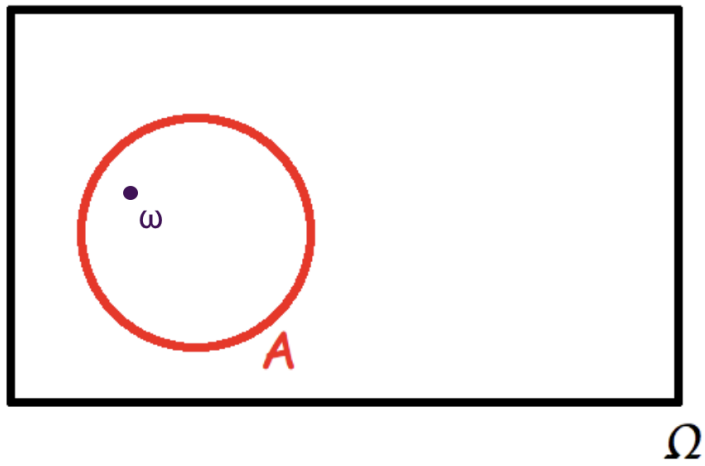
Et, idéalement, comme les axiomes d'Euclide, nous voulons une liste minimale mais suffisante d'axiomes permettant de générer une théorie riche.

Espace Fondamental et L'algèbre des Ensembles

On s'interesse au resultat d'un processus dont le résultat est incertain – une experience aléatoire.

Les résultats possibles, et toute affirmation les concernant doivent, être exprimés en termes de la théorie des ensembles.

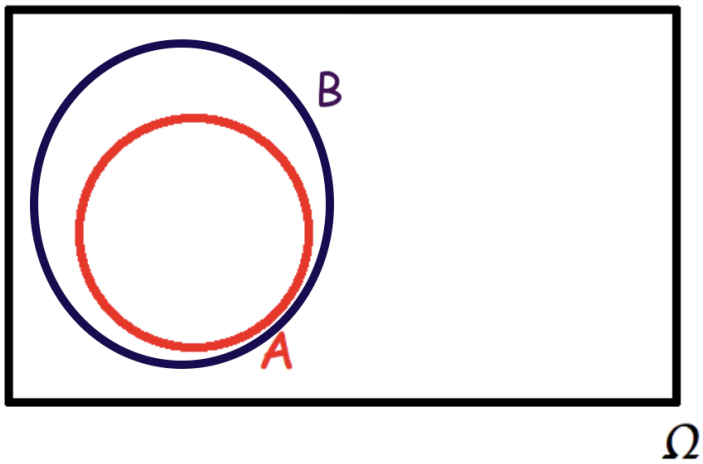
- Un résultat possible ω est appelé un événement élémentaire.
- L'ensemble de tous les résultats possibles Ω, est appelé l'espace fondemental.
- Un sous-ensemble $A \subseteq \Omega$ de Ω peut réunir plusieurs événements élémentaires.
- Un sous-ensemble $A \subseteq \Omega$ "est réalisé" (ou "se produit") lorsque le résultat de l'expérience appartient à A.



A **implique** $B \leftrightarrow A \subseteq B$ (inclusion)

- A est sous-ensemble de B, écrit $A \subseteq B$, lorsque $\omega \in A \implies \omega \in B$.
- Si $A \subseteq B$ et aussi $A \subseteq B$, alors les deux ensembles sont égaux, $A = B$.
- Si A s'est realisé, alors il est necessaire que B s'est realisé aussi.
- L'inclusion est transitive: si $A \subseteq B$ et $B \subseteq C$, alors $A \subseteq C$.
- Jet d'un dé: "obtenir deux" implique "obtenir un chiffre pair"

$$\{2\} \subseteq \{2, 4, 6\}$$



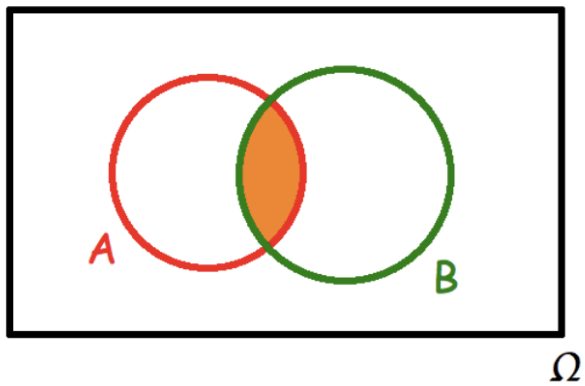
A **et** $B \leftrightarrow A \cap B$ (intersection)

- L'intersection des ensembles contient tous les événements élémentaires communs contenus dans les deux événements, et seulement ceux.
- Équivalemment, $\omega \in A \cap B$ si et seulement si $\omega \in A$ et $\omega \in B$,

$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ et } \omega \in B\}$$

- Deux ensembles A et B sont dits disjoints (ou incompatibles) si leur intersection est vide: $A \cap B = \emptyset$ (ils ne contiennent aucun élément commun).
- L'intersection est symétrique: $A \cap B = B \cap A$
- Jet d'un dé: "obtenir un chiffre pair" et "obtenir un chiffre premier"

$$\{2, 4, 6\} \cap \{2, 3, 5\} = \{2\}$$



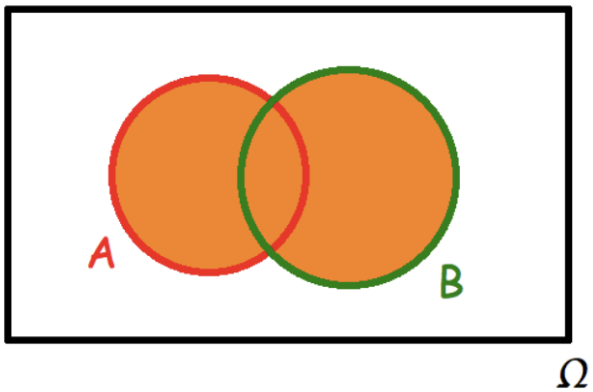
A **ou** $B \leftrightarrow A \cup B$ (union)

- L'union contient tous les événements élémentaires contenus dans les deux ensembles, tels qu'ils sont.
- Équivalemment, $\omega \in A \cup B$ si et seulement si $\omega \in A$ ou $\omega \in B$,

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ ou } \omega \in B\}$$

- $A \cup B = \emptyset \iff A = \emptyset \,\&\, B = \emptyset$.
- L'union est symétrique: $A \cup B = B \cup A$
- Jet d'un dé: "obtenir un chiffre pair" ou "obtenir un chiffre premier"

$$\{2,4,6\} \cup \{2,3,5\} = \{2,3,4,5,6\}$$



$A \cup B$

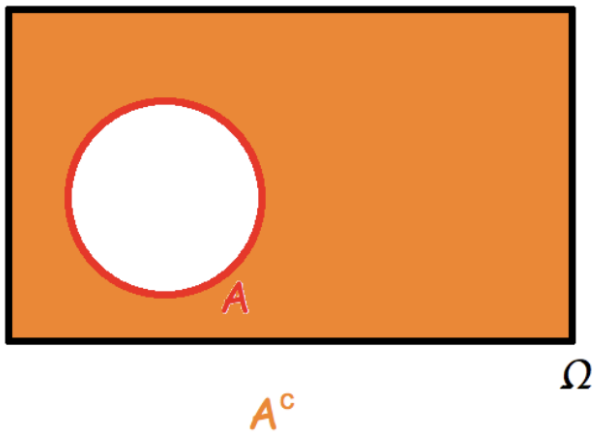
pas $A \leftrightarrow A^c$ (complement)

- Le complement de A, A^c, contient tous les événements elementaires de Ω qui ne sont pas contenus dans A, et seulement ceux-là,

$$A^c = \{\omega \in \Omega : \omega \notin A\}.$$

- Le complement de A sera *vide* (ou *impossible*), seulement si $A = \Omega$.
- Evidemment: $A \cup A^c = \Omega$, $A \cap A^c = \emptyset$
- Le complement reverse l'ordre: $A \subseteq B \iff A^c \supseteq B^c$.
- Jet d'un dé: pas "obtenir un chiffre pair"

$$\{2, 4, 6\}^c = \{1, 3, 5\}$$



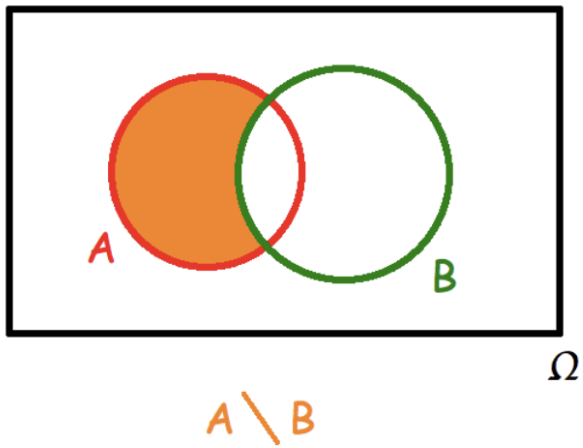
A **mais pas** $B \leftrightarrow A \setminus B = A \cap B^c$ (difference)

- La difference entre A et B, $A \setminus B$, contient tous les événements élémentaires contenus dans A, sauf ceux qui sont contenus aussi dans B,

$$A \setminus B = A \cap B^c$$

- La différence n'est pas symétrique en générale:
  $A \setminus B = A \cap B^c \neq B \cap A^c = B \setminus A$
- La difference $A \setminus B$ sera *vide*, seulement si $A \subseteq B$
- Jet d'un dé: "obtenir un chiffre pair" mais pas "obtenir un chiffre premier"

$$\{2, 4, 6\} \setminus \{2, 3, 5\} = \{4, 6\}$$



Associativité:

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

Distributivité:

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

Lois de De Morgan:

- $(A \cup B)^c = A^c \cap B^c$
- $(A \cap B)^c = A^c \cup B^c$

Lois de De Morgan générales:

- $\left(\bigcup_{i \geq 1} A_i\right)^c = \bigcap_{i \geq 1} A_i^c$
- $\left(\bigcap_{i \geq 1} A_i\right)^c = \bigcup_{i \geq 1} A_i^c$

(Notation: $\bigcup_{i \geq 1} A_i = A_1 \cup A_2 \cup \ldots$ et $\bigcap_{i \geq 1} A_i = A_1 \cap A_2 \cap \ldots$)

On vient de voir d'ensembles de la forme $A_1 \cup A_2 \cup \ldots$ et $A_1 \cap A_2 \cap \ldots$
Par associativité, on peut les comprendre de manière iterative:

- $\bigcup_{i\geq 1} A_i$ signifie: A_1 ou A_2 ou A_3...
- $\bigcap_{i\geq 1} A_i$ signifie: A_1 et A_2 et A_3...

Équivalemment,

- $\omega \in \bigcup_{i\geq 1} A_i$ si et seulement si $\exists\, i \geq 1 : \omega \in A_i$.
- $\omega \in \bigcap_{i\geq 1} A_i$ si et seulement si $\omega \in A_i$, $\forall i \geq 1$.

En général, ayant une suite $\{A_n\}$ monotone, on peut définir les operations limites:

- (suite décroissante) $A_{n+1} \subseteq A_n \implies \lim_{n\to\infty} A_n = \bigcap_{n\geq 1} A_n$
- (suite croissante)$A_{n+1} \supseteq A_n \implies \lim_{n\to\infty} A_n = \bigcup_{n\geq 1} A_n$

Pour une suites générale $\{A_n\}$ non-monotone, $C_n = \cap_{j\geq n} A_j$ est croissante en n et $D_n = \cup_{j\geq n} A_j$ est décroissante en n, alors on peut définir :
"$\{A_n\}$ finalement" $\leftrightarrow \liminf_n A_n = \cup_{n\geq 1} \cap_{j\geq n} A_j = \lim_{n\to\infty} C_n$

$$\liminf_n A_n = \{\omega \in \Omega : \exists n \geq 1 \text{ tel que } \omega \in A_j \,\forall\, j \geq n\}$$

$\liminf A_n$ se réalise $\iff$ apart une sous collection finie, tous les A_n se réalisent

"$\{A_n\}$ infiniment souvent" $\leftrightarrow \limsup_n A_n = \cap_{n\geq 1} \cup_{j\geq n} A_j = \lim_{n\to\infty} D_n$

$$\limsup_n A_n = \{\omega \in \Omega : \forall\, n \geq 1\, \exists j \geq n \text{ tel que } \omega \in A_j\}$$

$\limsup A_n$ se réalise $\iff$ tous les A_n dans une sous-collection infinie se réalisent

Example (Un singe tapant Shakespeare)

Un singe frappe au hasard les touches d'une machine à écrire, sans fin.

- L'espace fondementale est l'ensemble des suites infinies de chaînes de caractères A–Z.
- Chaque événement élémentaire est de la forme $\omega = (\omega_1, \omega_2, ...)$ avec $\omega_j \in \{A, ..., Z\}$.
- Définissons l'evenement que la $n$ème "treizaine" soit TOBEORNOTTOBE:

$$\begin{aligned} A_n = \{\omega \quad : \quad & \omega_{13(n-1)+1} = T, \omega_{13(n-1)+2} = O, \omega_{13(n-1)+3} = B, \omega_{13(n-1)+4} = E, \\ & \omega_{13(n-1)+5} = O, \omega_{13(n-1)+6} = R, \omega_{13(n-1)+7} = N, \\ & \omega_{13(n-1)+8} = O, \omega_{13(n-1)+9} = T, \omega_{13(n-1)+10} = T, \\ & \omega_{13(n-1)+11} = O, \omega_{13(n-1)+12} = B, \omega_{13(n-1)+13} = E\}. \end{aligned}$$

- $\liminf_n A_n \leftrightarrow$ "le singe finira par taper TOBEORNOTTOBETOBEORNOTTOBE... sans fin"
- $\limsup_n A_n \leftrightarrow$ "le singe tapera TOBEORNOTTOBE une infinité de fois"
- Évidemment, on peut remplacer TOBEORNOTTOBE par une chaîne de caractères finie de n'importe quelle longueur — par exemple l'ensemble des œuvres de Shakespeare.

Dans des situations simples, on peut parfois observer directement l'issue exacte d'une expérience aléatoire.

- Exemple : lancer d'un dé parfait $\rightarrow$ on observe un entier de 1 à 6.

Mais dans beaucoup de cas, cela n'est pas possible :

- L'information révélée peut être partielle.
- L'appareil de mesure peut avoir une résolution limitée.

Cela signifie que les "observables" possibles ne sont pas toujours tous les sous-ensembles de Ω (la puissance 2^{Ω}), mais plutôt une sous-collection $\mathcal{F} \subseteq 2^{\Omega}$.

(A est "observable" si, à l'issue de l'expérience, on peut savoir si A s'est réalisé)

Question : Toute collection $\mathcal{F}$ de sous-ensembles est-elle une "bonne" collection d'observables ?

Voici quelques propriétés naturelles que devraient respecter les observables :

- Ω est toujours observable — quelque chose s'est produit !
- Si on peut observer si A s'est réalisé, alors on peut aussi l'observer pour A^c.
- Si A et B sont observables, alors $A \cup B$ et $A \cap B$ le sont aussi.

Ces règles expriment une forme de stabilité algébrique : les observables devraient former une algèbre d'ensembles.

Mais on en veut souvent plus :

- si on peut observer une suite dénombrable d'événements $\{A_n\}_{n=1}^{\infty}$, on souhaite aussi pouvoir observer l'union $\bigcup_{n=1}^{\infty} A_n$, ou l'intersection $\bigcap_{n=1}^{\infty} A_n$.
- Cela nous permettra d'observer les liminf/limsup, qui permettent de formuler des propositions logiques non seulement *finies*, mais aussi plus élaborées — comme "à partir d'un certain rang", "une infinité de fois", etc.

Cela conduit à la notion de σ-algèbre — une collection stable par opérations dénombrables.

Definition (σ-algèbre)

Soit Ω un ensemble. Une σ-algèbre (ou tribu) $\mathcal{F} \subseteq 2^{\Omega}$ est une collection de sous-ensembles de Ω telle que :

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F} \implies A^c \in \mathcal{F}$ (stabilité par complémentaire)
3. $\{A_n\}_{n\geq 1} \subseteq \mathcal{F} \implies \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ (stabilité par réunion dénombrable)

Étant donnée une paire $(\Omega, \mathcal{F})$, on appelera les éléments ω de Ω des événements simples et les éléments A de $\mathcal{F}$ des événements.

Remarques:

- Par les règles de De Morgan, (3) implique qu'on a aussi la stabilité par intersection dénombrable ($\{A_n\}_{n\geq 1} \subseteq \mathcal{F} \implies \bigcap_{n=1}^{\infty} A_n \in \mathcal{F}$).
- Si on remplace (3) par $\boxed{A, B \in \mathcal{F} \implies A \cup B \in \mathcal{F}}$, alors $\mathcal{F}$ est appelé une algèbre sur Ω plutôt qu'une σ-algèbre.

Remarques supplémentaires:

- On peut définir plusieurs σ-algèbres différentes sur un même Ω.
- Parfois le choix de la σ-algèbre est imposé par ce que l'on est capable d'observer.
- Le choix peut dependre aussi de considerations mathématiques.
  - Si Ω est dénombrable, on prend souvent pour $\mathcal{F}$ l'ensemble de tous les sous-ensembles de Ω.
  - Mais si Ω est non dénombrable (par exemple $\mathbb{R}$), on ne peut pas faire ainsi — cela mène à des inconsistances mathématiques (hors du cadre de ce cours).

Example (Plusieurs σ-algèbres)

Je lance deux dés simultanement, un rouge et un vert.

- Quelle est la σ-algèbre la plus riche, disons $\mathcal{F}_1$?
- Et si je suis daltonien, quelle est ma σ-algèbre $\mathcal{F}_2$?
- J'informe mon colleague seulement du total. Quel est sa σ-algèbre $\mathcal{F}_3$?

Considérons l'espace $\Omega = \{(r, v) : r, v \in \{1, \ldots, 6\}\}$ (lancer de deux dés).

$r \setminus v$	$v=1$	$v=2$	$v=3$	$v=4$	$v=5$	$v=6$
$r=1$	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
$r=2$	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
$r=3$	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
$r=4$	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
$r=5$	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
$r=6$	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Cas (a) : information complète On observe l'issue exacte $\omega = (r, v)$. $\Rightarrow$ La tribu $\mathcal{F}_1$ est le plus riche possible :

$$\mathcal{F}_1 = 2^{\Omega}, \quad |\mathcal{F}_1| = 2^{36}$$

Cas (b) : perception symétrisée (daltonien) Impossible de distinguer (r, v) de (v, r) ; par exemple $(1, 2)$ et $(2, 1)$ sont confondus. $\Rightarrow$ Les observables sont de la forme

$$\{(i, j)\} \cup \{(j, i)\} \equiv \{(i, j), (j, i)\}, \text{ ou } \{(i, i)\}$$

Il y en a $6(6 + 1)/2 = 21$ classes, donc $|\mathcal{F}_2| = 2^{21}$.

Cas (c) : somme des dés seulement On connaît seulement $t = r + v$ (valeurs possibles de 2 à 12). $\Rightarrow$ Les observables sont de la forme $A_t := \{(i, j) : i + j = t\} = \bigcup_{i+j=t} \{(i, j)\}$, t=2,...,12. Observons les A_t sont les 11 "antidiagonales" : {(1,1)}, {(1,2),(2,1)}, {(1,3),(2,2),(3,1)},... . Alors $|\mathcal{F}_2| = 2^{11}$. Observez qu'on peut savoir la somme a partir de $\mathcal{F}_2$, mais on ne peut pas savoir le résultat symetrisé a partir de $\mathcal{F}_3$ (car une somme de 4, peut être 2+2 ou 1+3...)

On a : $\mathcal{F}_1 \supset \mathcal{F}_2 \supset \mathcal{F}_3$. Les σ-algèbres $\mathcal{F}_2$ et $\mathcal{F}_3$ représentent un appauvrissement progressif de $\mathcal{F}_1$ — seules des questions moins précises peuvent être posées ou résolues.

Espaces de Probabilité

Il existe de nombreuses propriétés souhaitables que l'on pourrait attendre d'une mesure de probabilité.

- Se comporter comme un volume (additivité sur les événements disjoints, et monotonie).
- Représenter la limite, à long terme, des fréquences relatives.
- ...

Certaines de ces propriétés sont si intuitives qu'elles ont même été proposées comme définitions de la probabilité — mais de telles tentatives ont historiquement échoué à produire une théorie robuste.

La contribution de Kolmogorov a été d'identifier l'*ensemble minimal d'axiomes* suffisant pour construire une théorie mathématique puissante et cohérente.

- Ces axiomes permettent de démontrer des résultats profonds comme la loi des grands nombres, en tant que *théorèmes*.

Un autre avantage de cette définition est son neutralité interprétative :

- Si vous croyez au hasard — tant mieux !
- Sinon, vous pouvez tout de même l'utiliser comme une mesure de l'incertitude.

A.N. Kolmogorov (1933), "Foundations of the Theory of Probability"

Axiomes de probabilités

Une fonction de probabilité sur une pair $(\Omega, \mathcal{F})$ composé d'un ensemble fondamental Ω et une σ-algèbre de sous-ensembles de Ω doit satisfaire:

1. Positivité : $\mathbb{P}(A) \geq 0$ pour tout $A \in \mathcal{F}$.
2. Événement certain : $\mathbb{P}(\Omega) = 1$.
3. σ-additivité : pour toute suite $\{A_n\}_{n\geq 1} \subset \mathcal{F}$ d'événements deux-à-deux disjoints[a],

$$\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) = \sum_{n\geq 1} \mathbb{P}(A_n)$$

[a] c'est à dire $A_i \cap A_j = \emptyset$ lorsque $i \neq j$.

Un triplet $(\Omega, \mathcal{F}, \mathbb{P})$ satisfaisant ces conditions est appelé un espace de probabilité ou une experience aléatoire.

Proposition (propriétés fondamentales)

Pour tout espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$

- Événement impossible : $\emptyset \in \mathcal{F}$, $\mathbb{P}(\emptyset) = 0$
- Additivité finie : si $A_i \cap A_j = \emptyset$ pour $i \neq j$, alors $\mathbb{P}(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mathbb{P}(A_i)$.
- Complémentaire : $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ pour tout $A \in \mathcal{F}$.
- Monotonie : si $A_1, A_2 \in \mathcal{F}$ satisfont $A_1 \subseteq A_2$, alors $\mathbb{P}(A_1) \leq \mathbb{P}(A_2)$
- Formules de éxclusion/inclusion :

$$\begin{aligned}
\mathbb{P}(A_1 \cup A_2) &= \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2) \\
\mathbb{P}(A_1 \cup A_2 \cup A_3) &= \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) - \mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_1 \cap A_3) - \mathbb{P}(A_2 \cap A_2) \\
&\qquad + \mathbb{P}\{A_1 \cap A_2 \cap A_3\}
\end{aligned}$$

$$\mathbb{P}\left(\cup_{i=1}^n A_i\right) = \sum_{g=1}^{n} (-1)^{g+1} \sum_{1 \leq i_1 < \ldots < i_g \leq g} \mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_g})$$

- σ-sous-addidivité: pour toute séquence $\{A_n\}$, $\mathbb{P}(\cup_{n=1}^\infty A_n) \leq \sum_{n=1}^\infty \mathbb{P}(A_n)$.
- Continuité monotone : pour toute séquence $\{A_n\}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\cup_{i=1}^n A_i\right) = \mathbb{P}\left(\cup_{i=1}^\infty A_i\right) \quad \& \quad \lim_{n\to\infty} \mathbb{P}\left(\cap_{i=1}^n A_i\right) = \mathbb{P}\left(\cap_{i=1}^\infty A_i\right)$$

Démonstration.

Événement impossible : $\emptyset = \Omega^c \in \mathcal{F}$ (car $\Omega \in \mathcal{F}$ et $\mathcal{F}$ est stable par complémentation). Comme $\emptyset \cup \emptyset \cup \ldots = \emptyset$, on applique la σ-additivité et positivité : $0 \leq \mathbb{P}(\emptyset) = \mathbb{P}(\emptyset) + \mathbb{P}(\emptyset) + \ldots \implies \mathbb{P}(\emptyset) = 0$.

Additivité finie : On élarge la famille $A_1, \ldots, A_n$ en posant $A_j = \emptyset$ pour $j > n$. Ainsi, on a $\mathbb{P}\left(\cup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mathbb{P}(A_j)$ mais pour $j > n$, $\mathbb{P}(A_j) = 0$, donc : $\mathbb{P}(A_1 \cup \cdots \cup A_n) = \sum_{j=1}^{m} \mathbb{P}(A_j)$.

Complémentaire : Comme $\Omega = A \cup A^c$ avec $A \cap A^c = \emptyset$, par additivité :

$$\mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c) = 1 \Rightarrow \mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$

Monotonie : Si $A_1 \subseteq A_2$, alors $A_2 = (A_2 \setminus A_1) \cup A_1$, avec $(A_2 \setminus A_1) \cap A_1 = \emptyset$. Donc :

$$\mathbb{P}(A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2 \setminus A_1) \geq \mathbb{P}(A_1)$$

Formules d'inclusion/exclusion : On écrit $A_1 \cup A_2 = (A_1 \setminus A_2) \cup (A_1 \cap A_2) \cup (A_2 \setminus A_1)$ En utilisant l'additivité finie : $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1 \setminus A_2) + \mathbb{P}(A_1 \cap A_2) + \mathbb{P}(A_2 \setminus A_1)$. Mais $\mathbb{P}(A_1) = \mathbb{P}(A_1 \setminus A_2) + \mathbb{P}(A_1 \cap A_2)$, et $\mathbb{P}(A_2) = \mathbb{P}(A_2 \setminus A_1) + \mathbb{P}(A_2 \cap A_1)$, d'ou découle le résultat pour $n = 2$.

Pour le cas $n = 3$, comme on a $A_1 \cup A_2 \cup A_3 = A_1 \cup (A_2 \cup A_3)$,

$$\begin{aligned}\mathbb{P}(A_1 \cup A_2 \cup A_3) &= \mathbb{P}(A_1) + \mathbb{P}(A_2 \cup A_3) - \mathbb{P}(A_1 \cap (A_2 \cup A_3)) \\ &= \mathbb{P}(A_1) + \mathbb{P}(A_2 \cup A_3) - \mathbb{P}((A_1 \cap A_2) \cup (A_1 \cap A_3)) \\ &= \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) - \mathbb{P}(A_1 \cap A_2) \\ &\quad - \mathbb{P}(A_1 \cap A_3) - \mathbb{P}(A_2 \cap A_3) + \mathbb{P}((A_1 \cap A_2) \cap (A_1 \cap A_3))\end{aligned}$$

et observons que le dernier terme est $\mathbb{P}(A_1 \cap A_2 \cap A_3)$. La formule generale est prouvé en iterant cet argument.

σ-sous-addidivité: Écrivons $\bigcup_{i=1}^{\infty} A_i = A_1 \cup \bigcup_{i=2}^{\infty} A_i = A_1 \cup B$, et remarquons que, via monotonie, on a

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathbb{P}(A_1 \cup B) \leq \mathbb{P}(A_1) + \mathbb{P}(B) \leq \cdots$$

Continuité monotone : Soit $C_n = \bigcup_{i=1}^{n} A_i$. On remarque que $C_n \subset C_{n+1}$ pour tout n, et donc $(C_{i+1} \setminus C_i) \cap (C_{j+1} \setminus C_j) = \emptyset$ si $i \neq j$ (faire un dessin). Ainsi, les ensembles $C_{n+1} \setminus C_n$ sont deux à deux disjoints et en plus $\mathbb{P}[C_j \setminus C_{j-1}] = \mathbb{P}(C_j) - \mathbb{P}(C_{j-1})$ car la sequence C_n est croissante.

$$\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) = \mathbb{P}(C_\infty) = \mathbb{P}(C_1) + \sum_{i=2}^{\infty} \mathbb{P}(C_i \setminus C_{i-1})$$

$$= \mathbb{P}(C_1) + \sum_{i=2}^{\infty} \{\mathbb{P}(C_i) - \mathbb{P}(C_{i-1})\}$$

$$= \lim_{n\to\infty} \left[\mathbb{P}(C_1) + \sum_{i=2}^{n} \{\mathbb{P}(C_i) - \mathbb{P}(C_{i-1})\}\right]$$

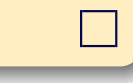
$$= \lim_{n\to\infty} \mathbb{P}(C_n) = \lim_{n\to\infty} \mathbb{P}\left(\cup_{i=1}^{n} A_i\right)$$

où nous avons utilisé le fait que $\mathbb{P}(C_n)$ est croissante et majorée, donc converge. Pour la deuxième partie, observez que par de Morgan:

$$\lim_{n\to\infty} \mathbb{P}[\cap_{j=1}^{n} A_j] = \lim_{n\to\infty} (1 - \mathbb{P}[(\cap_{j=1}^{n} A_j)^c]) = 1 - \lim_{n\to\infty} \mathbb{P}[\cup_{j=1}^{n} A_j^c] = 1 - \mathbb{P}[\cup_{j=1}^{\infty} A_j^c].$$

où nous avons utilisé la première partie à la fin. Utilisons de Moivre à nouveau :

$$1 - \mathbb{P}[\cup_{j=1}^{\infty} A_j^c] = 1 - (1 - \mathbb{P}[(\cup_{j=1}^{\infty} A_j^c)^c]) = \mathbb{P}[\cap_{j=1}^{\infty} A_j]$$



Soient $\Omega = \{\omega_1, ..., \omega_r\}$ et $\mathcal{F} = 2^{|\Omega|}$. Est-ce qu'on peut

characteriser les mesures de probabilité possibles sur $(\Omega, \mathcal{F})$?

Posons $\mathbb{P}(\{\omega_i\}) \equiv \mathbb{P}(\omega_i) \equiv p_i$. Alors $p_1, \ldots, p_r$ sont des réels tels que

$$p_1, ..., p_r \geq 0 \quad \& \quad p_1 + \cdots + p_r = 1.$$

La donnée des nombres p_i détermine la probabilité $\mathbb{P}(G)$ de n'importe quel événement $G \in \mathcal{F}$. En effet, si $G = \{\omega_{i_1}, \ldots, \omega_{i_k}\}$, alors par l'additivité finie :

$$P(G) = P(\{\omega_{i_1}\} \cup \cdots \cup \{\omega_{i_k}\}) = P(\omega_{i_1}) + \cdots + P(\omega_{i_k}) = p_{i_1} + \cdots + p_{i_k}.$$

Par conséquent :

$$P(G) = \sum_{j\,:\,\omega_j \in G} p_j.$$

Réciproquement, il est possible de se donner des nombres $p_1, \ldots, p_r$ vérifiant $p_i \geq 0$, $\sum p_i = 1$, de poser $P(\omega_i) = p_i$, puis de définir $\mathbb{P}(G)$ par la formule ci-dessus lorsque $G \neq \emptyset$, et sinon définir $\mathbb{P}(\emptyset) = 0$.

Exercice : Toute fonction $\mathbb{P} : \mathcal{F} \to [0, 1]$ ainsi définie, vérifie les axiomes.

Soit $\Omega = \{\omega_1, \ldots, \omega_n\}$ un espace de cardinal fini n, et soit $\mathcal{F} = 2^\Omega$ l'ensemble de toutes les parties.

Une probabilité est dite uniforme si :

$$\mathbb{P}(\{\omega_i\}) = \frac{1}{n} \quad \text{pour tout } i = 1, \ldots, n.$$

Dans ce cas, pour tout événement $A \subseteq \Omega$, on a :

$$\mathbb{P}(A) = \frac{\text{nombre d'issues favorables}}{\text{nombre total d'issues}} = \frac{|A|}{|\Omega|}$$

Ce cadre est adapté aux modèles discrets élémentaires
(dés, cartes, tirages sans remise, etc.)

Une telle définition est évidemment compatible avec les axiomes, selon notre discussion dans le cas générale.

Pause pensée : Lorsque l'espace Ω est fini, on peut toujours en nommer les éléments comme $\{1, 2, \ldots, |\Omega|\}$. Cependant, il est souvent plus commode de les décrire de manière plus verbale ou même symbolique, à condition de rester clair.

Example (deux pièces)

On lance deux pièces de monnaie équitables – quelle est la probabilité que de $G =$ "deux résultats identiques" ?

$$\Omega = \{(\mathrm{P},\mathrm{P}),(\mathrm{P},\mathrm{F}),(\mathrm{F},\mathrm{P}),(\mathrm{F},\mathrm{F})\}$$

et $\#\Omega = 4$. Alors $G = \{(\mathrm{P},\mathrm{P}),(\mathrm{F},\mathrm{F})\}$, donc $|G| = 2$. En admettant que les quatre résultats élémentaires sont équiprobables, $\mathbb{P}(G) = \frac{|G|}{|\Omega|} = \frac{2}{4} = \frac{1}{2}$.

Example (tirage d'une carte)

On tire une carte au hasard d'un jeu standard de 52 cartes.

- L'univers Ω contient 52 issues.
- La probabilité uniforme donne $\mathbb{P}(\{\omega\}) = \frac{1}{52}$ pour toute carte ω.

Exemples d'événements :

- A : "la carte est un pique ♠" $\Rightarrow \mathbb{P}(A) = \frac{13}{52} = \frac{1}{4}$
- B : "la carte est un as A" $\Rightarrow \mathbb{P}(B) = \frac{4}{52} = \frac{1}{13}$
- C : "la carte est rouge" $\Rightarrow \mathbb{P}(C) = \frac{26}{52} = \frac{1}{2}$

Example (paradoxe de Monty Hall)

Ce paradoxe célèbre vient d'un jeu télévisé américain *Let's Make a Deal*, animé par Monty Hall dans les années 1960. Il a suscité un débat intense lorsqu'il a été popularisé par Marilyn vos Savant dans la presse dans les années 1990.

On nous présente trois portes : derrière l'une se cache une voiture, derrière les deux autres, une chèvre. Nous choisissons une porte au hasard. L'animateur, qui sait ce qu'il y a derrière chaque porte, toujours ouvre l'une des deux autres portes qui cache une chèvre. Il vous propose alors de changer de porte.

Doit-on changer ? Pour répondre, considérons les 3 cas pour notre choix initiale :

- *Dans 1 cas sur 3*, on a choisi la voiture. Changer = perdre.
- *Dans 2 cas sur 3*, on a choisi une chèvre. Changer = gagner.

Changer donne double la probabilité de gagner – Il est toujours beaucoup plus préférable de changer de porte !

Surprenant?! Même le grand mathématicien Paul Erdös s'est montré sceptique — il n'a été convaincu qu'en considérant une version avec *un million de portes*, dont une seule cache une voiture – et l'animateur ouvre 999'998 portes avec chèvres.

Example (Problème des anniversaires)

Quelle est la probabilité qu'il n'y ait pas de coïncidences d'anniversaires dans une classe de n étudiants (en éliminant les hypothèses d'années bissextiles, des jumeaux, etc.) ? On a

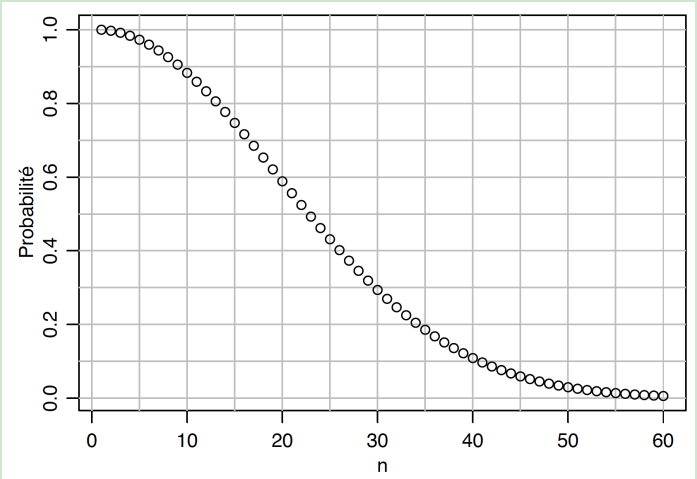
$$\Omega = \{(i_1, \ldots, i_n) : i_1, \ldots, i_n \in \{1, \ldots, 365\}\}, \quad \text{avec} \quad |\Omega| = 365^n,$$

et

$$A_n = \{(i_1, \ldots, i_n) : \text{tous les } i_j \text{ sont distincts}\}.$$

Ainsi,

$$|A_n| = 365 \times \cdots \times (365 - n + 1) = \frac{365!}{(365-n)!} \quad \text{et} \quad \mathbb{P}(A_n) = \frac{|A_n|}{|\Omega|} = \frac{365!}{(365-n)!365^n}.$$



Souvent, on doit compter les nombres d'éléments d'ensembles finis de grande cardinalité, pour lesquels une simple énumération des éléments est impossible.

Comment ? Deux principes de base :

- *addition :*ayant n pantalons et m t-shitrs, j'ai un total de $n+m$ vêtements.
- *multiplication :* ayant n pantalons et m t-shirts, il y a nm tenues.

D'après les deux principle, voiçi quelques formules utiles.

Permutations

Une permutation de n objets distincts $a_1, \ldots, a_n$ est un *arrangement ordonné*, sans répétition, de ces n objets. D'après le principe de multiplication,

$$\#\text{ de permutations possibles de } n \text{ objets} = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1 = n!$$

Remarque : Par convention, on pose $0! = 1$. La valeur de $n!$ augmente très rapidement avec n. On a la propriété asymptotique, appelée *formule de Stirling* :

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n,$$

où $a_n \sim b_n$ signifie que $\lim_{n\to\infty}(a_n/b_n) = 1$.

Arrangements

Un arrangement est une permutation de k objets pris parmi n objets distincts ($k \leq n$). Les objets sont donc choisis *sans répétition* et de manière *ordonnée*.

Puisqu'il y a n choix possibles pour le premier objet, puis $n-1$ choix pour le deuxième, $n-2$ pour le troisième, etc., le nombre d'arrangements est

$$A_{n,k} = n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}$$

Parfois, l'ordre n'est pas important – seulement la *collection* nous interesse:

Combinaisons

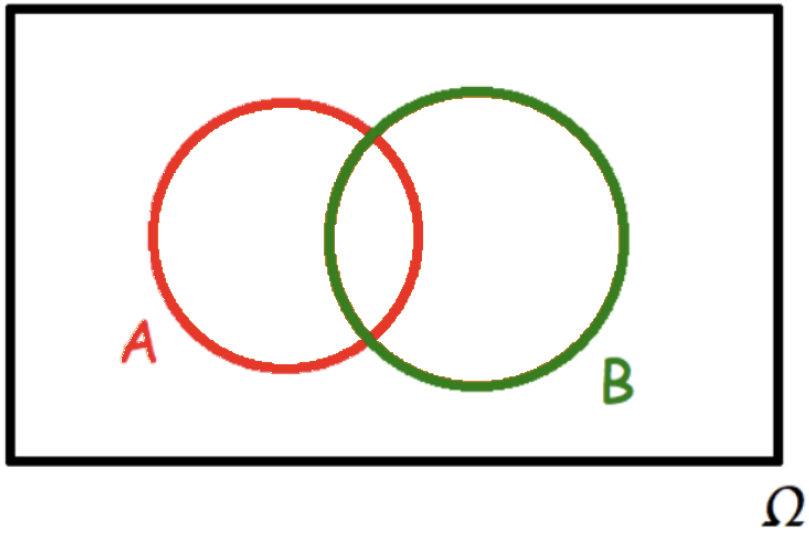
Une combinaison de k éléments pris parmi n éléments (distincts) est un *sous-ensemble à k éléments* de cet ensemble de n éléments. Le nombre de ces sous-ensembles est noté $C_{n,k}$ ou $\binom{n}{k}$ ($=$ le *coefficient binomial*).

En observant que à chaque sous-ensemble correspondent $k!$ arrangements, nous obtenons que $A_{n,k} = C_{n,k} \cdot k!$, et alors

$$C_{n,k} \equiv \binom{n}{k} = \frac{A_{n,k}}{k!} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!}$$

Probabilité Conditionnelle et Independence

- Soient $A, B \in \mathcal{F}$ deux événements.
- La probabilité que A se réalise est $\mathbb{P}(A)$.
- Et s'il nous est révélé que B s'est produit, cela change-t-il quelque chose ?
- Dans un sens, l'espace fondamental a evolué: $\Omega \longrightarrow B$
- Certains événements sont plus possibles, autres ont une probabilité différente.



Probabilité conditionnelle

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ une espace de probabilité, et $B \in \mathcal{F}$ un événement. Si $\mathbb{P}(B) > 0$, on définit :

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \qquad \forall A \in \mathcal{F}.$$

la "probabilité (conditionnelle) de A sachant (que) B (est réalisé)"

Remarque : *Si* $\mathbb{P}(B) = 0$, *on adopte la convention* $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B)$, *des deux côtés on a la valeur zéro. Ainsi*

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) = \mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid B^c)\mathbb{P}(B^c)$$

même si $\mathbb{P}(B) = 0$ *ou* $\mathbb{P}(B^c) = 0$.

Remarque : Attention, en général $\mathbb{P}(A \mid B) \neq 1 - \mathbb{P}(A \mid B^c)$

Example

Une urne contient 6 boules rouges et 5 noires. On tire deux boules sans remise. Quelle est la probabilité que la deuxième soit noire, sachant que la première est rouge ?

$$\mathbb{P}(H \mid G) = \frac{6 \times 5}{11 \times 10} \Big/ \frac{6 \times 10}{11 \times 10} = \frac{5}{10} = \frac{1}{2}.$$

Théorème

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ une espace de probabilité, et $B \in \mathcal{F}$ un événement. Si $\mathbb{P}(B) > 0$, alors $\mathbb{Q}(A) = \mathbb{P}(A \mid B)$ est une mésure de probabilité sur $(\Omega, \mathcal{F})$.

Preuve.

Il suffit de vérifier les axiomes. Si $A \in \mathcal{F}$, alors

$$\mathbb{Q}(A) = \mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \in [0, 1],$$

car $A \cap B \subset B$ et donc $\mathbb{P}(A \cap B) \leq \mathbb{P}(B)$. De même,

$$\mathbb{Q}(\Omega) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1,$$

et enfin, pour des $\{A_n\} \subset \mathcal{F}$ avec $A_i \cap A_j$, $i \neq j$,

$$\mathbb{Q}\left(\bigcup_{i=1}^{\infty} A_i\right) = \frac{\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i \cap B\right)}{\mathbb{P}(B)} = \frac{\mathbb{P}\left(\bigcup_{i=1}^{\infty} (A_i \cap B)\right)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \mathbb{Q}(A_i),$$

en utilisant les propriétés de $\mathbb{P}(\cdot)$ et le fait que si $A_1, A_2, \ldots$ sont deux à deux disjoints, alors les $A_j \cap B, \ldots$ le sont aussi. □

Munis du dernier résultat, one peut établir une relation élégante entre le conditionnement sur une intersection $B_1 \cap B_2$ et les conditionnements successifs :

- Soient $B_1, B_2 \in \mathcal{F}$ tels que $\mathbb{P}(B_1 \cap B_2) > 0$ et $\mathbb{Q}_i(\cdot) = \mathbb{P}(\cdot \mid B_i)$, $i = 1, 2$.
- Par la définition :

$$\mathbb{P}(A \mid B_1 \cap B_2) = \frac{\mathbb{P}(A \cap B_1 \cap B_2)}{\mathbb{P}(B_1 \cap B_2)} = \frac{\mathbb{P}(A \cap B_2 \mid B_1)\mathbb{P}(B_1)}{\mathbb{P}(B_2 \mid B_1)\mathbb{P}(B_1)}.$$

- Alors on peut aussi écrire

$$\mathbb{P}(A \mid B_1 \cap B_2) = \frac{\mathbb{P}(A \cap B_2 \mid B_1)}{\mathbb{P}(B_2 \mid B_1)} = \frac{\mathbb{Q}_1(A \cap B_2)}{\mathbb{Q}_1(B_2)} = \mathbb{Q}_1(A \mid B_2).$$

Informellement :

$$\mathbb{P}(A \mid B_1 \cap B_2) = \text{“}\mathbb{P}(A \mid B_1 \mid B_2)\text{”}$$

Interprétation :

- Conditionner sur $B_1 \cap B_2$ revient à conditionner d'abord sur B_1, puis sur B_2.
- L'ordre peut être inversé : $\mathbb{P}(A \mid B_2 \cap B_1) = \text{“}\mathbb{P}(A \mid B_2 \mid B_1)\text{”}$.

Notation : Parfois nous écrivons $\mathbb{P}(A \mid B_1, B_2)$ au lieu de $\mathbb{P}(A \mid B_1 \cap B_2)$.

Souvent, il est plus facile de calculer une probabilité en la décomposant à l'aide du conditionnement. Deux telles décompositions sont données par :

1. La décomposition de prévision, qui simplifie le calcul des probabilités d'intersections, en de "prévisions iteratives".
2. La formule des probabilités totales, qui découpe un événement en plusieurs morceaux plus maniables.

Proposition (Décomposition de prévision)

Soient $A_1, \dots, A_n \in \mathcal{F}$ des événements d'un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. Alors

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)$$

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_2 \cap A_1)$$

$$\vdots$$

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_2 \cap A_1) \cdots \mathbb{P}(A_n \mid A_{n-1} \cap \cdots \cap A_1).$$

Exercice : Vérifiez l'énonce en itérant sa premiere conclusion.

Pause pensée : Interprétez l'énonce via conditionnement successif.

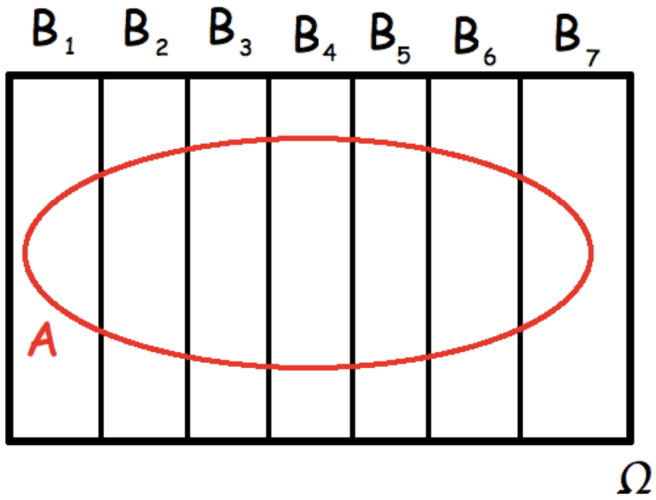
Proposition (Formule des probabilités totales)

Soit $\{B_1, \ldots, B_n\} \subset \mathcal{F}$ une partition de Ω, c'est à dire:

- $B_i \cap B_j = \emptyset \; \forall \; i \neq j$.
- $B_1 \cup \ldots \cup B_n = \Omega$.

Alors pour tout événement $A \in \mathcal{F}$,

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\, \mathbb{P}(B_i).$$



Remarques:

- La preuve est effectivement dans l'énoncé.
- La notion de partition (et la conclusion du théorème) reste valable lorsqu'on prend $n = \infty$ (alors $\{B_i\}_{i=1}^{\infty}$ est une séquence).

Example (Boule rouge au second tirage)

Une urne contient n_1 boules rouges, n_2 boules noires et n_3 boules bleues. Soit $n = n_1 + n_2 + n_3$. On effectue deux tirages sans remise. Quelle est la probabilité de l'événement $A =$"la deuxième boule tirée est rouge" ?
Soit $\{A_1, A_2, A_3\}$ une partition de Ω définie comme suit:

$$A_1 = \text{“la première boule tirée est rouge”},$$
$$A_2 = \text{“la première boule tirée est noire”},$$
$$A_3 = \text{“la première boule tirée est bleue”}.$$

Par conséquent, en utilisant la formule des probabilités totales,

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \mid A_1)\mathbb{P}(A_1) + \mathbb{P}(A \mid A_2)\mathbb{P}(A_2) + \mathbb{P}(A \mid A_3)\mathbb{P}(A_3) \\ &= \frac{n_1 - 1}{n-1} \cdot \frac{n_1}{n} + \frac{n_1}{n-1} \cdot \frac{n_2}{n} + \frac{n_1}{n-1} \cdot \frac{n_3}{n} = \frac{n_1(n_1 - 1 + n_2 + n_3)}{n(n-1)} = \frac{n_1}{n}.\end{aligned}$$

Il est intéressant d'observer que $\frac{n_1}{n}$ est aussi la probabilité que la première boule tirée soit rouge ! En fait, que les tirages soient effectués avec remise ou sans remise ne change pas la probabilité de l'événement "la deuxième boule tirée est rouge". La même propriété est valable si on tire n boules et qu'on s'intéresse à la probabilité de l'événement "la $i^{\text{ème}}$ boule tirée est rouge" $(i \leq n)$.

Observons d'abord que, sauf dans de cas exceptionnelles, on aura

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

Cela fait beaucoup de sens, en considérant des exemples specifiques:

- $\mathbb{P}[\text{gagner la loterie} \mid \text{avoir un ticket}] \neq \mathbb{P}[\text{avoir un ticket} \mid \text{gagner la loterie}]$
- $\mathbb{P}[\text{être hospitalisé} \mid \text{avoir le COVID}] \neq \mathbb{P}[\text{avoir le COVID} \mid \text{être hospitalisé}]$

La formule de Bayes, lie $\mathbb{P}(A \mid B)$ à $\mathbb{P}(B|A)$. C'est un outil simple mais très important, tant en terms probabilistes qu'en termes de logique.

Théorème (Bayes)

Soient deux événements $A, B \in \mathcal{F}$ avec $\mathbb{P}(A) > 0$, alors :

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\,\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \mid B)\,\mathbb{P}(B)}{\mathbb{P}(A \mid B)\,\mathbb{P}(B) + \mathbb{P}(A \mid B^c)\,\mathbb{P}(B^c)}.$$

Preuve.

Découle de la définition et de la formule des probabilités totales. □

Example (Probabilité conditionnelle d'hospitalisation)

En été 2022 aux États-Unis :

- 28.1% des hospitalisés étaient non-vaccinés, 71.9% "vaccinés" (série primaire ou rappel). C'est à dire, avec les événements d'intérêt définis de manière évidente,

$$\mathbb{P}(V \mid H) = 0.719 \qquad \& \qquad \mathbb{P}(V^c \mid H) = 0.281.$$

- Mais on s'intéresse plutôt a l'inverse: la probabilité d'être hospitalisé, sachant notre status de vaccination.
- 80% de la population était vaccinée, donc 20% non-vaccinée.
- Taux d'hospitalisation dans la population générale : $\mathbb{P}(H) = 0.01$.

Nous utilisons la formule de Bayes,

$$\mathbb{P}(H \mid V) = \frac{\mathbb{P}(V \mid H) \cdot \mathbb{P}(H)}{\mathbb{P}(V)} = \frac{0.719 \cdot 0.01}{0.80} \approx 0.00899 \text{ (soit } 0.899\%)$$

$$\mathbb{P}(H \mid V^c) = \frac{\mathbb{P}(V^c \mid H) \cdot \mathbb{P}(H)}{\mathbb{P}(V^c)} = \frac{0.281 \cdot 0.01}{0.20} \approx 0.01405 \text{ (soit } 1.405\%)$$

- Si on ne comprends pas bien la difference entre $\mathbb{P}(V \mid H)$ et $\mathbb{P}(H \mid V)$, ça peut donner de très différentes interpretations de la situation.

Distinguons maintenant entre personnes "âgées" (≥ 60 ans) vs "jeunes" (< 60 ans)

- 20% de la population a ≥ 60 ans ; parmi les hospitalisations, 80% sont âgée.
- Parmi les hospitalisés âgées : 85% vaccinés, 15% non-vaccinés.
- Dans la sous-population âgée : 90% vaccinés, 10% non-vaccinés.

$$\mathbb{P}(H \mid 60+) = \frac{\mathbb{P}(60+ \mid H)\mathbb{P}(H)}{\mathbb{P}(60+)} = \frac{0.008}{0.2} = 0.04$$

$$\mathbb{P}(H \mid V, 60+) = \frac{\mathbb{P}(V \mid H, 60+) \cdot \mathbb{P}(H \mid 60+)}{\mathbb{P}(V \mid 60+)} = \frac{0.85 \cdot 0.04}{0.9} \approx 0.0378$$

$$\mathbb{P}(H \mid V^c, 60+) = \frac{\mathbb{P}(V^c \mid H, 60+) \cdot \mathbb{P}(H \mid 60+)}{\mathbb{P}(V^c \mid 60+)} = \frac{0.15 \cdot 0.04}{0.1} \approx 0.06$$

- 80% de la population a < 60 ans ; parmi les hospitalisations, 20% sont jeunes.
- Parmi les hospitalisés jeunes : 50% vaccinés, 50% non-vaccinés.
- Dans la sous-population jeune : 70% vaccinés, 30% non-vaccinés.

$$\mathbb{P}(H \mid <60) = \frac{\mathbb{P}(<60 \mid H)\mathbb{P}(H)}{\mathbb{P}(<60)} = \frac{0.002}{0.8} = 0.0025$$

$$\mathbb{P}(H \mid V, <60) = \frac{\mathbb{P}(V \mid H, <60) \cdot \mathbb{P}(H \mid <60)}{\mathbb{P}(V \mid <60)} = \frac{0.50 \cdot 0.0025}{0.70} \approx 0.00179$$

$$\mathbb{P}(H \mid V^c, <60) = \frac{\mathbb{P}(V^c \mid H, <60) \cdot \mathbb{P}(H \mid <60)}{\mathbb{P}(V^c \mid <60)} = \frac{0.50 \cdot 0.0025}{0.30} \approx 0.00417$$

En resumé :

Groupe	$\mathbb{P}(\text{hospitalisation} \mid \text{groupe})$
Vaccinés, ≥ 60 ans	3.78%
Non-vaccinés, ≥ 60 ans	6.00%
Vaccinés, < 60 ans	0.179%
Non-vaccinés, < 60 ans	0.417%

En utilisant la formule des probabilités totales, on obtient aussi:

Corollaire (formule de Bayes généralisé)

Soit $\{B_i\} \subset \mathcal{F}$ une partition de Ω et soit $A \in \mathcal{F}$ tel que $\mathbb{P}(A) > 0$. Alors,

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i)}{\sum_i \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}.$$

Remarque : Observez que $\{B, B^c\}$ est toujours une partition de Ω, alors la formule de Bayes simple est un cas particulier de la formule générale lorsqu'il y a deux elements dans la partition.

Intuition. Dire que "*A et B sont indépendants*" signifie que la réalisation de l'un des deux n'affecte pas la réalisation de l'autre. C'est-à-dire que

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

donc la connaissance de la réalisation de B laisse $\mathbb{P}(A)$ inchangée.

Définition (Indépendence)

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité. Deux événements $A, B \in \mathcal{F}$ sont indépendants (on écrit $A \perp\!\!\!\perp B$) si

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

Conformément à notre intuition, cela implique que

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

et par symétrie $\mathbb{P}(B \mid A) = \mathbb{P}(B)$.

Remarque : Ne pas confonder indépendance avec incompatibilité! Par exemple,

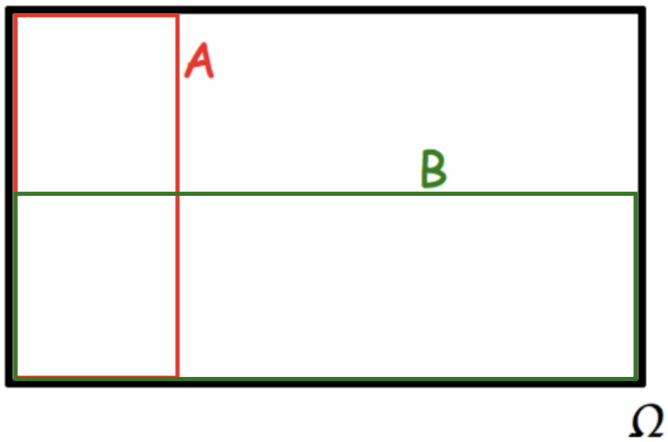
$$\begin{aligned} A \cap B = \emptyset \quad &\Longrightarrow \quad A, B \text{ incompatibles/disjoints} \\ &\Longrightarrow \quad \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B). \end{aligned}$$

A, B disjoints avec $\mathbb{P}(A), \mathbb{P}(B) > 0$ implique

$$\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0, \quad \text{mais} \quad \mathbb{P}(A) \times \mathbb{P}(B) \neq 0,$$

donc A et B sont dépendants – sachant A, on est sur que B n'est plus possible !

L'indépendence est plus subtile – ça implique une "conservation de proportions"



- Supposons que $\mathbb{P}$ représente l'aire.
- Alors $A \perp\!\!\!\perp B$ implique que le rapport de surface de A et de Ω, et le même que le rapport de surface de $A \cap B$ et de B.
- En terms de proportions : la proportion des A dans la population générale est la même que dans la sous-population des B.

Example (Deux enfants dans une famille)

Une famille a deux enfants.

1. On sait que le **premier** est un garçon. Quelle est la probabilité que le **second** soit aussi un garçon ?
2. On sait que **l'un des deux** est un garçon. Quelle est la probabilité que **l'autre** soit un garçon ?

L'espace fondamental est $\Omega = \{\text{GG}, \text{GF}, \text{FG}, \text{FF}\}$. Définissons les événements :

- $A_1 = \{\text{GG}, \text{GF}\}$: "le premier enfant est un garçon"
- $A_2 = \{\text{GG}, \text{FG}\}$: "le second enfant est un garçon"

(1) On cherche $\mathbb{P}(A_2 \mid A_1)$. On a : $\mathbb{P}(A_2 \mid A_1) = \dfrac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_1)} = \dfrac{1/4}{1/2} = 1/2 = \mathbb{P}(A_2)$.

Donc B_2 et B_1 sont indépendants.

(2) L'événement "au moins un est un garçon" est $C = A_1 \cup A_2 = \{\text{GG}, \text{GF}, \text{FG}\}$. Attention : dire "que **l'autre** soit un garçon" revient implicitement à supposer qu'il y a déjà un garçon. L'événement est donc équivalent à "les deux sont des garçons", soit $D = \{\text{GG}\}$. Alors,

$$\mathbb{P}(D \mid C) = \frac{\mathbb{P}(D \cap C)}{\mathbb{P}(C)} = \frac{\mathbb{P}(D)}{\mathbb{P}(C)} = \frac{1/4}{3/4} = 1/3 \neq \mathbb{P}(D).$$

Donc D et C ne sont pas indépendants.

Remarquons l'importance d'un langage précis : dans le cas (a), on sait qu'un enfant spécifique est un garçon, tandis que dans le cas (b), on sait seulement que l'un des deux enfants est un garçon. Ces informations, bien que similaires en apparence, ne sont pas équivalentes !

Indépendence : 2-à-2, mutuelle, conditionelle

1. Les événements $A_1, \ldots, A_n$ sont indépendants 2-à-2 si

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j), \quad 1 \leq i < j \leq n.$$

2. Les événements $A_1, \ldots, A_n$ sont (mutuellement) indépendants si pour tout sous-ensemble d'indices $F \subset \{1, \ldots, n\}$, on a :

$$\mathbb{P}\left(\cap_{i \in F} A_i\right) = \prod_{i \in F} \mathbb{P}(A_i).$$

3. Les événements $A_1, \ldots, A_n$ sont conditionnellement indépendants sachant B si pour tout sous-ensemble d'indices $F \subset \{1, \ldots, n\}$, on a :

$$\mathbb{P}\left(\bigcap_{i \in F} A_i \,\middle|\, B\right) = \prod_{i \in F} \mathbb{P}(A_i \mid B).$$

Remarques :

- L'indépendance est une idée clé qui simplifie beaucoup des calculs de probabilité. En pratique, il est essentiel de vérifier si les événements sont indépendants, car une hypothèse erronée d'indépendance peut modifier grandement le résultat.

- Les événements indépendants 2-à-2 ne sont pas forcément mutuellement indépendants.

- L'indépendance mutuelle entraîne l'indépendance conditionnelle, mais l'inverse est vrai seulement quand $B = \Omega$.

- L'indépendance conditionnelle est essentielle pour distinguer une dépendance directe d'une dépendance indirecte. Considérons les événements

$$A_1 = \text{“être attaqué par un requin”}$$
$$A_2 = \text{“avoir consommé une glace”}$$
$$B = \text{“c'est l'été”},$$

  et réfléchissons à l'indépendance ou à l'indépendance conditionnelle (sachant B) entre A_1 et A_2.

Lemme (indépendance et complémentaires)

Si $\{A_1, \ldots, A_n\}$ sont mutuellement indépendants, alors toute famille obtenue en remplaçant certains (ou tous) les A_i par leurs complémentaires l'est également.

Proof.

Considérons deux événements independents $A_1 = A$ et $A_2 = B$. On a

$$\begin{aligned}\mathbb{P}(A^c \cap B^c) \;=\; 1 - \mathbb{P}(A \cup B) \quad &= \quad 1 - (\mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)) \\ &= \quad 1 - \Big(\mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A)\,\mathbb{P}(B)\Big) \\ &= \quad (1 - \mathbb{P}(A))(1 - \mathbb{P}(B)) = \mathbb{P}(A^c)\,\mathbb{P}(B^c).\end{aligned}$$

De plus,

$$\begin{aligned}\mathbb{P}(A \cap B^c) = \mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(A \cap B) \quad &= \quad \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) \\ &= \quad \mathbb{P}(A)\big(1 - \mathbb{P}(B)\big) \\ &= \quad \mathbb{P}(A)\,\mathbb{P}(B^c).\end{aligned}$$

Le cas général est un peu plus laborieux et fait intervenir la formule d'inclusion-exclusion générale □

The epistemological value of probability theory is based on the fact that chance phenomena, considered collectively and on a grand scale, create non-random regularity.

A.N. Kolmogorov

- De manière remarquable, en théorie des probabilités, "l'ordre" apparaît souvent "à la limite".
- Cette limite peut correspondre à une grande échelle, un long terme, etc.
- Les événements limites tombent souvent dans les extrêmes : probabilité 0/1.

Les lemmes de Borel–Cantelli sont un premier exemplede ce phénomène. Nous allons les utiliser pour répondre à une question que nous avons posée à propos des longues séquences dans des lancers de pièces (et du singe qui tape Shakespeare).

Rappel :

"A_n finalement" $\leftrightarrow \liminf_n A_n$ (liminf)

$$\liminf_n A_n = \bigcup_{n\geq 1}\bigcap_{j\geq n} A_j = \lim_{n\to\infty}\bigcap_{j\geq n} A_j = \{\omega\in\Omega : \exists n\geq 1 \text{ tel que } \omega\in A_j \text{ pour tout } j\geq n\}$$

"A_n infiniment souvent" $\leftrightarrow \limsup_n A_n$ (limsup)

$$\limsup_n A_n = \bigcap_{n\geq 1}\bigcup_{j\geq n} A_j = \lim_{n\to\infty}\bigcup_{j\geq n} A_j = \{\omega\in\Omega : \forall n\geq 1, \exists j\geq n \text{ tel que } \omega\in A_j\}$$

Lemme (Borel-Cantelli)

Soient $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $\{A_n\} \subset \mathcal{F}$ une suite d'événements.

1. Premier lemme : sans aucune supposition supplémentaire sur les $\{A_n\}_{n\geq 1}$,

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \implies \mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 0.$$

2. Deuxième lemme : à condition que les $\{A_n\}_{n\geq 1}$ sont indépendants,

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty \implies \mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 1.$$

Observez que, par de Moivre,

$$\left(\limsup_{n\to\infty} A_n\right)^c = \left(\cap_{n\geq 1} \cup_{j\geq n} A_j\right)^c = \cup_{n\geq 1} \cap_{j\geq n} A_j^c = \liminf_{n\to\infty} A_n^c.$$

Attention : si $\sum_{n=1}^{N} \mathbb{P}(A_n)$ converge, alors $\sum_{n=1}^{N} \mathbb{P}(A_n^c) = N - \sum_{n=1}^{N} \mathbb{P}(A_n)$ diverge. Mais la divergence de $\sum_{n=1}^{\infty} \mathbb{P}(A_n)$ n'implique pas la convergence de $\sum_{n=1}^{\infty} \mathbb{P}(A_n^c)$!

Preuve du premier lemme.

Par définition

$$\limsup A_n = \cap_{i=1}^{\infty} \cup_{j=i}^{\infty} A_j \subset \cup_{j=i}^{\infty} A_k, \quad \forall\, i \geq 1.$$

Alors par monotonicité et σ-sous-additivité,

$$\mathbb{P}\left(\limsup A_n\right) \leq \mathbb{P}\left(\bigcup_{j=i}^{\infty} A_j\right) \leq \sum_{j=i}^{\infty} \mathbb{P}(A_j) \quad \forall\, i \geq 1.$$

Mais comme la série $\sum_{j=1}^{\infty} \mathbb{P}(A_j)$ converge, on a $\sum_{j=i}^{\infty} \mathbb{P}(A_k) \overset{i\to\infty}{\to} 0$. La partie à gauche ne dépend pas de l'indice i, alors en prenant la limite à gauche et à droite,

$$\mathbb{P}\left(\limsup A_n\right) \leq \lim_{i\to\infty} \sum_{j=i}^{\infty} \mathbb{P}(A_j) = 0.$$

□

Preuve du deuxième lemme.

Remarquons $(\limsup A_n)^c = \liminf A_n^c = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c = \lim_{n\to\infty} \bigcap_{k=n}^{\infty} A_k^c$ car $C_n = \bigcap_{k=n}^{\infty} A_k^c$ est une suite croissante en n. Donc :

$$\mathbb{P}\big[\,(\limsup A_n)^c\,\big] = \mathbb{P}\left(\lim_{n\to\infty} \cap_{k=n}^{\infty} A_k^c\right) = \lim_{n\to\infty} \mathbb{P}\left(\cap_{k=n}^{\infty} A_k^c\right)$$

par continuité monotone de $\mathbb{P}$. Maintenant $\mathbb{P}\left(\cap_{k=n}^{m} A_k^c\right) = \prod_{k=n}^{m}(1-\mathbb{P}(A_k))$ par l'indépendance. Utilisant l'inégalité $\log(1-x) \leq -x \quad$ pour , $0 < x < 1$, on a

$$\prod_{k=n}^{m}(1-\mathbb{P}(A_k)) = \exp\log\left[\prod_{k=n}^{m}(1-\mathbb{P}(A_k))\right] = \exp\left[\sum_{k=n}^{m}\log(1-\mathbb{P}(A_k))\right] \leq \exp\left[-\sum_{k=n}^{m}\mathbb{P}(A_k)\right]$$

Mais comme $\sum_{k=1}^{\infty}\mathbb{P}(A_k) = \infty$, on a $\sum_{k=n}^{m}\mathbb{P}(A_k) \overset{m\to\infty}{\to} \infty$, alors la borne à droite converge vers 0 lorsque $m \to \infty$. Alors par continuité monotone de $\mathbb{P}$,

$$\mathbb{P}\left(\bigcap_{k=n}^{\infty} A_k^c\right) = \mathbb{P}\left(\lim_{m\to\infty}\bigcap_{k=n}^{m} A_k^c\right) = \lim_{m\to\infty}\mathbb{P}\left(\bigcap_{k=n}^{m} A_k^c\right) \leq \lim_{m\to\infty}\exp\left[-\sum_{k=n}^{m}\mathbb{P}(A_k)\right] = 0.$$

et ceci pour tout $n \geq 1$. La suite $p_n = \mathbb{P}\left(\cap_{k=n}^{\infty} A_k^c\right)$ est donc nulle, et sa limite est nulle aussi. □

Example (Séquences de piles de plus en plus longues)

On considère une suite infinie de lancers de pièces équilibrées et indépendantes (pile ou face). Pour chaque $n \in \mathbb{N}$, définissons l'événement :

$$A_n = \{\text{"Les lancers de } n \text{ à } 2n \text{ donnent tous pile"}\}$$

Alors :

$$\mathbb{P}(A_n) = \left(\frac{1}{2}\right)^n \quad \Rightarrow \quad \sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$$

Conclusion (Borel–Cantelli 1) :

$$\mathbb{P}\left(A_n \text{ se produit infiniment souvent}\right) = 0$$

Interprétation :

- On ne verra presque jamais une infinité de séquences de plus en plus longues entièrement en pile.
- Autrement dit, avec probabilité 1 il existe un rang N tel que pour tout $n \geq N$, les lancers de n à $2n$ ne sont jamais tous pile.

Example (Treizaines de "pile" et Borel–Cantelli 2)

On lance une pièce idéale indéfiniment, les lancers étant independents avec $\mathbb{P}(\mathrm{P}) = \mathbb{P}(\mathrm{F}) = 1/2$.

- L'espace fondamental est l'ensemble des suites infinies $\omega = (\omega_1, \omega_2, \dots)$ avec $\omega_j \in \{\mathrm{P}, \mathrm{F}\}$.
- Chaque événement élémentaire est une trajectoire de lancers ω.
- Définissons l'événement que la n-ième treizaine (bloc non chevauchant de longueur 13) soit composée uniquement de piles :

$$\begin{aligned} A_n = \{\omega \quad : \quad & \omega_{13(n-1)+1} = \mathrm{P}, \omega_{13(n-1)+2} = \mathrm{P}, \omega_{13(n-1)+3} = \mathrm{P}, \omega_{13(n-1)+4} = \mathrm{P}, \\ & \omega_{13(n-1)+5} = \mathrm{P}, \omega_{13(n-1)+6} = \mathrm{P}, \omega_{13(n-1)+7} = \mathrm{P}, \\ & \omega_{13(n-1)+8} = \mathrm{P}, \omega_{13(n-1)+9} = \mathrm{P}, \omega_{13(n-1)+10} = \mathrm{P}, \\ & \omega_{13(n-1)+11} = \mathrm{P}, \omega_{13(n-1)+12} = \mathrm{P}, \omega_{13(n-1)+13} = \mathrm{P}\}. \end{aligned}$$

- Alors $\mathbb{P}(A_n) = 2^{-13}$ pour tout n, et grâce au découpage en blocs non chevauchants, les (A_n) sont indépendants.
- $\limsup_{n\to\infty} A_n \leftrightarrow$ "il y a une infinité de treizaines toutes piles."
- Comme $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \sum_{n=1}^{\infty} 2^{-13} = +\infty$ et que les (A_n) sont indépendants, le lemme de Borel–Cantelli 2 donne $\mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 1$.
- Comparer avec le singe tapant Shakespeare : là, on veut 13 caractères exacts, et la probabilité qu'un bloc de 13 corresponde est 26^{-13} (ce qui ne change pas grand-chose). Évidemment, on peut choisir n'importe quel $k < \infty$ au lieu de 13 (par exemple l'ensemble des œuvres de Shakespeare), mais pas des propositions de plus en plus longues (voir l'exemple précédent).

Variables aléatoires et fonctions de répartition

Variables aléatoires : résumés numériques de l'issue d'une expérience aléatoire.

Elles nous permettent de ne pas nous soucier de la structure précise d'un résultat $\omega \in \Omega$. On peut se concentrer sur l'image d'une variable aléatoire, plutôt que sur Ω lui-même.

- Soit $(\Omega, \mathcal{F})$ un espace fondamental muni d'une σ-algèbre, et considérons une fonction réelle
$$X : \Omega \to \mathbb{R}, \qquad \omega \mapsto X(\omega).$$
- Écrivons $\{a \leq X \leq b\}$ pour désigner l'ensemble $\{\omega \in \Omega : a \leq X(\omega) \leq b\}$.
- Pour pouvoir attribuer une probabilité à l'ensemble $\{a \leq X \leq b\}$, celui-ci doit être un événement valide (observable), c'est-à-dire un élément de $\mathcal{F}$.
- Si $\{X \leq x\} \in \mathcal{F}$ pour tout réel $x \in \mathbb{R}$, alors on dit que X est une fonction mesurable, ou plus familièrement, une variable aléatoire.
- Si $\{X \leq x\} \in \mathcal{F}$ pour tout $x \in \mathbb{R}$, alors forcemment tout ensemble A obtenu à partir d'opérations dénombrables sur des événements de la forme $\{X \leq x_n\}$, $n \geq 1$, appartient également à $\mathcal{F}$.

Remarque : Si Ω possède une notion de distance, et que $\mathcal{F}$ est engendrée par des "boules" $\{\omega \in \Omega : d(\omega, \omega_0) \leq \varepsilon\}$, alors toute application continue $\Omega \to \mathbb{R}$ est également mesurable.

Example (Variable aléatoire indicatrice)

Soit $G \in \mathcal{F}$. On définit $1_G : \Omega \to \mathbb{R}$ par

$$1_G(\omega) = \begin{cases} 1 & \text{si } \omega \in G, \\ 0 & \text{sinon.} \end{cases}$$

La fonction 1_G *indique* par sa valeur si l'événement G se réalise ou pas. C'est bien une v.a., car

$$\{1_G \leq x\} = \begin{cases} \varnothing & \text{si } x < 0, \\ G^c & \text{si } 0 \leq x < 1, \\ \Omega & \text{si } x \geq 1. \end{cases}$$

Example (3 lancers de pièce)

On lance une pièce 3 fois. L'univers est donc $\Omega = \{0,1\}^3$, chaque $\omega = (\omega_1, \omega_2, \omega_3)$ représentant les résultats des 3 lancers. Soit $X_i = 1_{A_i}$ l'indicatrice de l'événement $A_i =$ "le lancer i donne pile", et définissons:

$$Y = X_1 + X_2 + X_3$$

c'est-à-dire le *nombre total de piles* obtenues.

- Observons qu'il s'agit bel et bien d'une variable aléatoire :

$$\begin{aligned}
\{Y = 0\} &= \{(0,0,0)\} \in \mathcal{F}, \\
\{Y = 1\} &= \{(1,0,0),(0,1,0),(0,0,1)\} \in \mathcal{F}, \\
\{Y = 2\} &= \{(1,1,0),(1,0,1),(0,1,1)\} \in \mathcal{F}, \\
\{Y = 3\} &= \{(1,1,1)\} \in \mathcal{F}.
\end{aligned}$$

- D'où il s'ensuit que $\{Y \leq x\} \in \mathcal{F}\ \forall x \in \mathbb{R}$, car on peut exprimer cet ensemble à l'aide d'opérations finies sur les ensembles ci-dessus.

Remarque. Bien que ω contienne la séquence complète des 3 lancers, $X(\omega)$ ne retient que le nombre total de piles. Ainsi, X ne transporte pas toute l'information de ω.

- Il s'ensuit que, si une mesure de probabilité $\mathbb{P}$ est définie sur les événements de Ω, alors il suffit de connaitre $\mathbb{P}(X \leq x)$ pour tout $x \in \mathbb{R}$, afin de connaitre le comportement aléatoire de X de manière complète.

- Pour cette raison, l'application réelle définie par $x \mapsto \mathbb{P}(X \leq x)$ joue un rôle central, et s'appelle la fonction de répartition de X ou la loi de de X,

$$\boxed{F_X : \mathbb{R} \to [0,1], \quad F_X(x) := \mathbb{P}[X \leq x], \quad x \in \mathbb{R}}$$

- À partir de F_X, on peut calculer toute probabilité de la forme $\mathbb{P}[X \in B]$, où B résulte d'opérations dénombrables ensemblistes portant sur des intervalles[1]

Par exemple, en utilisant la continuité monotone, on peut calculer

$$\mathbb{P}(X \in [a,b]) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} \{X \leq b\} \setminus \{X \leq a - \tfrac{1}{n}\}\right) = \lim_{n\to\infty} \mathbb{P}\big(a - \tfrac{1}{n} < X \leq b\big)$$

$$= F_X(b) - \lim_{n\to\infty} F_X\big(a - \tfrac{1}{n}\big) = F_X(b) - \lim_{a_n \uparrow a} F_X(a_n) = F_X(b) - F_X(a-).$$

[1] la σ-algèbre des sous-ensembles de $\mathbb{R}$ de cette forme est appelée la σ-algèbre borélienne de $\mathbb{R}$, dennotée par $\mathcal{B}(\mathbb{R})$; par abréviation, on parle d'ensembles boréliens. La loi de X peut-être vu comme une mesure de probabilité en soi même, defini sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Plus généralement, on peut montrer (exercice) :

Proposition (Probabilités à partir de la fonction de repartition)

Soient a et b deux nombres tels que $a \leq b$. Si F_X désigne la fonction de répartition (distribution function) de la variable aléatoire X, alors :

$$\begin{aligned}
\mathbb{P}\{X < a\} &= F_X(a-) \\
\mathbb{P}\{X > a\} &= 1 - F_X(a). \\
\mathbb{P}\{a < X \leq b\} &= F_X(b) - F_X(a). \\
\mathbb{P}\{a \leq X \leq b\} &= F_X(b) - F_X(a-). \\
\mathbb{P}\{a < X < b\} &= F_X(b-) - F_X(a), \quad \text{pour } a < b. \\
\mathbb{P}\{X = a\} &= F_X(a) - F_X(a-).
\end{aligned}$$

(Évidemment, si F_X est continue, $F_X(x) = F_X(x-)$, et les formules se simplifient.)

- Une fonction de répartition constitue une grande simplification : il suffit de prescrire une fonction réelle pour définir un modèle probabiliste pour X, sans se soucier outre mesure de l'espace de probabilité sous-jacent.
- Mais, bien entendu, n'importe quelle fonction réelle ne convient pas. Quels sont les propriétés qu'une fonction quelconque F doit satisfaire pour qu'elle soit une fonction de répartition valable?

Théorème (Characterisation de fonctions de répartition)

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $X : \Omega \to \mathbb{R}$ une variable aléatoire. Sa fonction de répartition F_X satisfait :

1. F_X est non-décroissante, ainsi $F_X(x) \leq F_X(y)$ pour $x \leq y$
2. $\lim_{x \to -\infty} F_X(x) = 0$ et $\lim_{x \to +\infty} F_X(x) = 1$
3. F_X est continue à droite, ainsi

$$\lim_{t \downarrow 0} F_X(x + t) = F_X(x), \quad x \in \mathbb{R}.$$

Réciproquement, pour toute fonction $F : \mathbb{R} \to \mathbb{R}$ qui satisfait (1)–(3), il existe un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et une variable aléatoire $X : \Omega \to \mathbb{R}$, tels que

$$F(x) = \mathbb{P}[X \leq x].$$

Nous ne démontrerons que l'énoncé direct. La réciproque est beaucoup plus profonde et nécessite des outils qui dépassent le cadre de ce cours.

(en particulier, la construction de la mesure de Lebesgue)

Preuve

(1) Non-décroissance. Si $y \geq x$, on peut écrire

$$F_X(y) = \mathbb{P}(X \leq y) = \mathbb{P}(X \leq x) + \underbrace{\mathbb{P}(x < X \leq y)}_{\geq 0} \geq \mathbb{P}(X \leq x).$$

(2) Limites. Soit $A_n = \{X \leq -n\}$. Alors $\{A_n\}_{n\in\mathbb{N}}$ est décroissante et $\lim_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} A_n = \emptyset$, car $X(\omega) \in \mathbb{R}$ pour tout ω (il n'y a pas de masse à $-\infty$, impossible d'être moindre que tout entier). Par continuité monotone,

$$0 = \mathbb{P}(\cap_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} F_X(-n).$$

Comme F_X est non décroissante, on en déduit $\lim_{n\to\infty} F_X(x_n) = 0$ pour toute suite $x_n \to -\infty$. L'autre limite est établi de façon similaire.
(d) Continuité à droite. Pour toute sequence $t_n \downarrow 0$, on a

$$F_X(x + t_n) = \mathbb{P}(X \leq x + t_n) = \mathbb{P}(X \leq x) + \mathbb{P}(x < X \leq x + t_n).$$

Le premier terme vaut $F_X(x)$. Puisque $A_n = (x, x + t_n]$ est une sequence décroissante qui se "rétrécit" vers $\emptyset$, et $B \mapsto \mathbb{P}(X \in B)$ est une mesure de probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, l'autre terme $\to 0$ par continuité monotone. □

Proposition (Points de discontinuité)

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité, et F_X la fonction de répartition d'une variable aléatoire $X : \Omega \to \mathbb{R}$. Alors l'ensemble des points de discontinuité de F_X est dénombrable, et égal à l'ensemble $D_F = \{x \in \mathbb{R} : \mathbb{P}(X = x) > 0\}$.

Preuve.

Notons D l'ensemble des discontinuités de la fonction F et D_n l'ensemble des discontinuités de F d'amplitude supérieure ou égale à $\frac{1}{n}$, c'est-à-dire

$$D = \{x \in \mathbb{R} : F(x) - F(x-) > 0\} \quad \text{et} \quad D_n = \left\{x \in \mathbb{R} : F(x) - F(x-) > \tfrac{1}{n}\right\}.$$

Alors $D = \bigcup_{n\in\mathbb{N}} D_n$. Montrons par l'absurde que $|D_n| < n$. Supposons donc que $|D_n| \geq n$. Choisissons alors $x_1 < \cdots < x_n \in D_n$. D'abord, par non-décroissance,

$$\sum_{i=1}^{n} \Big(F(x_i) - F(x_i-)\Big) = F(x_n) + \underbrace{\Big(-F(x_n-) + F(x_{n-1})\Big)}_{\leq 0} + \cdots + \underbrace{\Big(-F(x_2-) + F(x_1)\Big)}_{\leq 0} - F(x_1-).$$

$$\implies \sum_{i=1}^{n} \Big(F(x_i) - F(x_i-)\Big) \leq F(x_n) - F(x_1-) \leq F(+\infty) - F(-\infty) = 1.$$

Ensuite,

$$\sum_{i=1}^{n} \big(F(x_i) - F(x_i-)\big) > \sum_{i=1}^{n} \tfrac{1}{n} = \tfrac{n}{n} = 1.$$

Nous obtenons une contradiction, ce qui montre que $|D_n| < n$ pour tout n. Ainsi, l'ensemble $D = \bigcup_{n\in\mathbb{N}} D_n$ est une réunion dénombrable d'ensembles finis, et il est donc dénombrable.

Pour la deuxième partie, observons que

$$D_F = \{x : \mathbb{P}(X = x) > 0\} = \{x \in \mathbb{R} : F_X(x) > F_X(x-)\} = D.$$

C'est donc l'ensemble des discontinuités à gauche de F, et alors l'ensemble des discontinuités (comme F_X est continue à gauche). □

- Mais pourquoi une telle obsession pour les points de discontinuité ?
- Parce qu'ils vont nous permettre de classer les types possibles de variables aléatoires en trois catégories (dont deux majeures).

Soit D_F l'ensemble des discontinuités d'une fonction de repartition F_X.

- Si $D_F = \emptyset$ on appelle F_X une fonction de repartition continue (et X une v.a. continue)
- Si $\mathbb{P}(X \in D_F) = 1$ on appelle F_X une fonction de repartition discrète (et X une v.a. discrète)

Théorème (Classification des fonctions de répartition)

Toute fonction de répartition est une combinaison convexe d'une fonction de répartition discrète et d'une fonction de répartition continue.

Remarques :

- Autrement dit, toute fonction de répartition est soit continue, soit discrète, soit un mélange des deux.
- Vu le résultat du théorème il est habituel d'étudier séparément les v.a. discrètes et les v.a. continues, ce que nous allons faire dans la suite.
- Pour vérifier qu'une v.a. X est discrète, il suffit de montrer qu'il existe un sous-ensemble dénombrable $D \subset \mathbb{R}$ tel que $\mathbb{P}\{X \in D\} = 1$. Dans ce cas, on aura nécessairement $D_F \subset D$.

Le support $\mathcal{X} \equiv \text{supp}\{X\}$ d'une variable aléatoire X est défini comme

$$\begin{aligned} \mathcal{X} \equiv \text{supp}\{X\} &:= \{x \in \mathbb{R} : \mathbb{P}[|X - x| < \epsilon] > 0, \forall \epsilon > 0\} \\ &= \{x \in \mathbb{R} : F_X(b-) - F_X(a) > 0, \forall\, a < x < b\} \end{aligned}$$

Intuitivement, c'est la région de $\mathbb{R}$ que X peut "toucher".

C'est l'ensemble fermé $F \subseteq R$ le plus petit tel que $\mathbb{P}[X \in F] = 1$.

Exemples typiques que nous allons rencontrer :

1. Variables aléatoires continues, avec support :
   - L'axe réel entier, $\mathbb{R}$
   - Un demi-axe, par exemple $\mathbb{R}_+ = [0, +\infty)$
   - Un intervalle, par exemple $[0, 1]$.
2. Variables aléatoires discrètes, avec support :
   - Les entiers $\mathbb{Z}$
   - Les entiers non négatifs $\mathbb{Z}_+ = \{0, 1, 2, 3, \ldots\}$
   - L'ensemble $\{0, 1, 2, \ldots, n\}$ pour un $n \in \mathbb{N}$.
3. Variables aléatoires mixtes, avec support :
   - Un entier réuni à un intervalle, par exemple $\{0\} \cup [a, b]$, ou à un demi-axe, par exemple $\{0\} \cup [a, +\infty)$, pour $a \geq 0$.

Example

- Nombre d'appels téléphoniques durant un jour : $\mathcal{X} = \{0, 1, \ldots\}$.
- Nombre de piles pour n parties de pile ou face : $\mathcal{X} = \{0, 1, \ldots, n\}$.
- Temps d'attente du bus en minutes: $\mathcal{X} = [0, 15]$
- Temps d'attente avant l'émission d'une particule par un noyau instable : $\mathcal{X} = \mathbb{R}_+$.
- Quantité de pluie demain: $\mathcal{X} = \mathbb{R}_+$.
- Montant d'une indemnisation d'assurance, nul si aucun sinistre n'a lieu ou si le sinistre n'excède pas la franchise $a > 0$, et sinon supérieur ou égal à $a > 0$, $\mathcal{X} = \{0\} \cup [a, +\infty)$.
- Nombre de lancers nécessaires jusqu'à obtenir la première pile : $\mathcal{X} = \{1, 2, 3, \ldots\}$.
- Proportion d'humidité de l'air mesurée demain : $\mathcal{X} = [0, 1]$.
- Quantité de pluie demain : $\mathcal{X} = \{0\} \cup (0, +\infty) = [0, \infty)$.

- Chaque fois que nous cherchons à énoncer une affirmation probabiliste qui "cerne la position de X", cela revient à examiner des accroissements de F_X.
- Plus nous voulons localiser précisément nos affirmations probabilistes, plus il nous faut raffiner ces accroissements.
- Ces accroissements, finis dans le cas discret ou infinitésimaux dans le cas continu, conduisent naturellement aux fonctions de masse et de densité.

Proposition

Toute fonction de répartition F_X discrète admet une fonction de masse f_X,

$$f_X(x) = \lim_{\epsilon \downarrow 0} \big(F(x+\epsilon/2) - F(x-\epsilon/2)\big) \equiv F_X(x+) - F_X(x-) \equiv \mathbb{P}(X = x).$$

Il s'ensuit que toute fonction de masse satisfait :

1. $f_X(u) \geq 0$
2. $\sum_{u \in D_X} f(u) = 1$
3. $F_X(x) = \sum_{u \in D_X : u \leq x} f_X(u)$

Théorème (théorème de différentiation de Lebesgue, cas monotone)

Toute fonction de répartition F_X continue admet une fonction de densité f_X, telle que

$$f_X(u) = \lim_{\epsilon \downarrow 0} \left(\frac{F(x+\epsilon/2) - F(x-\epsilon/2)}{\epsilon} \right)$$

pour presque tout $u \in \mathbb{R}$.

Il s'ensuit que :

- Toute fonction de répartition F_X continue est même dérivable en presque tout $x \in \mathbb{R}$, avec dérivée $F'_X(x) = f_X(x)$.
- Comme $F'_X(x) \neq \underbrace{F_X(x) - F_X(x-)}_{=0} = \mathbb{P}(X = x)$, il s'agit d'une densité de probabilité, et non d'une probabilité. En fait $f_X : \mathbb{R} \to [0, +\infty)$!
- Souvent (et pour nos bésoins) F_X est en fait dérivable partout, et l'on a

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du \quad \text{et donc} \quad \int_{-\infty}^{+\infty} f_X(u)\, du = \lim_{x \to \infty} F_X(x) = 1.$$

(en réalité la formule ci-dessus est toujours valable, mais il faut une notion plus générale d'intégrale que l'intégrale de Riemann classique)

- Ainsi nous avons plutôt la correspondance "$\mathbb{P}(X \in (x, x+dx)) \approx f_X(x)\, dx$".

Cas discrèt vs cas continu

	Cas Discrèt	**Cas Continu**
Support $\mathcal{X}$	dénombrable	contient un intervalle (a, b), $a < b$.
f_X	fonction de masse, sans unité $f_X(x) \in [0,1], \ \forall\, x$ $\sum_{x \in \mathbb{R}} f_X(x) = 1$	fonction de densité, unité $[x]^{-1}$ $f_X(x) \in \mathbb{R}_+, \ \forall\, x$ $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$
$F_X(a) = \mathbb{P}(X \leq a)$	$\sum_{x \leq a} f_X(x)$	$\int_{-\infty}^{a} f_X(x)\, dx$
$\mathbb{P}(X \in B),\ B \in \mathcal{B}(\mathbb{R})$	$\sum_{x \in B} f_X(x)$	$\int_{B} f_X(x)\, dx$
$\mathbb{P}(a < X \leq b)$	$\sum_{\{x : a < x \leq b\}} f_X(x)$	$\int_{a}^{b} f_X(x)\, dx$
$\mathbb{P}(X = a)$	$f_X(a) \geq 0$	$\int_{a}^{a} f_X(x)\, dx = 0$

(encore une fois, dans le cas continu, nous utilisons une notion plus générale d'intégrale quand F_X n'est pas partout dérivable; en ligne 4, en particulier, on peut s'en faire une idée en remarquant que tout $A \in \mathcal{B}(\mathbb{R})$ peut être approché par des unions et intersections dénombrables d'intervalles)

Lois fondamentaux

Nous allons maintenant passer en revue certaines lois de probabilité de base, certaines discrètes et d'autres continues.

Afin de spécifier un modèle de probabilité discret, il suffit de :

1. Spécifier le support $\mathcal{X}$, c'est-à-dire un ensemble discret

$$\mathcal{X} = \{x : \mathbb{P}[X = x] > 0\}.$$

2. La valeur de la fonction de masse $f_X(x)$, en tant que fonction de $x \in \mathcal{X}$.

Nous nous concentrerons sur des lois telles que $\mathcal{X} \subseteq \mathbb{Z}$.

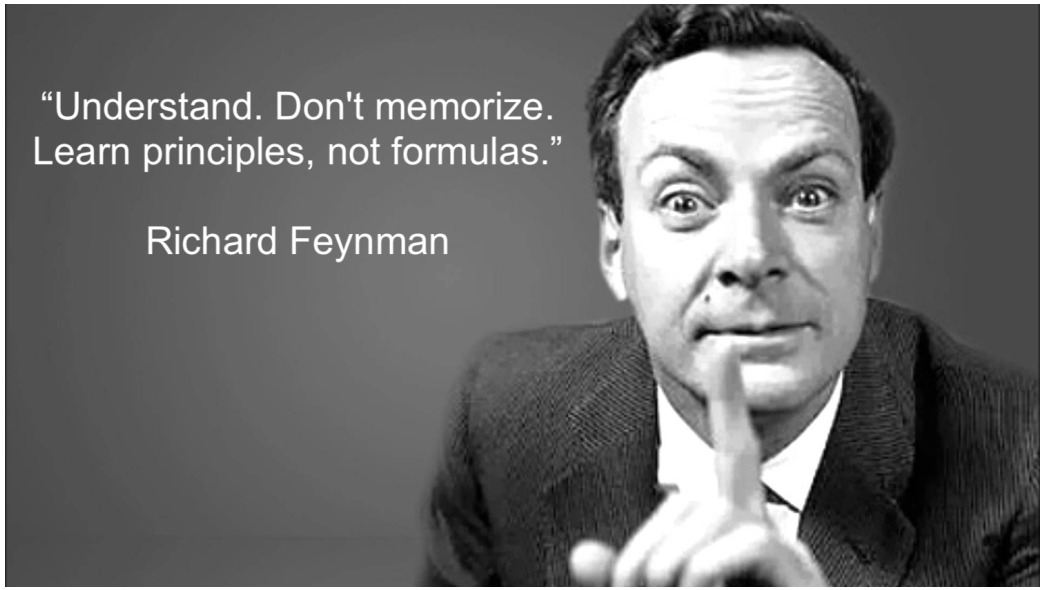
Afin de spécifier un modèle de probabilité continu, il suffit de :

1. Définir la fonction de densité de probabilité $f_X(x)$, en tant que fonction de $x \in \mathcal{X}$.
2. Spécifier le support $\mathcal{X}$, si celui-ci n'est pas a priori clair d'après la définition de la densité.

Nous nous concentrerons sur des lois telles que $\mathcal{X} = \mathbb{R}$ ou $\mathcal{X} = [0, \infty)$.

Dans chaque cas, il est important de:

- réfléchir à la manière dont elles peuvent apparaître à partir de principes premiers
- essayer de comprendre quelles caractéristiques la formule de $f_X(\cdot)$ reflète.



Avertissement

Certains symboles ne sont pas encore définis ! (il s'agit de $\mathbb{E}$, var, $M_X(\cdot)$)
Ne vous en souciez pas pour l'instant. Une fois que nous aurons introduit les moments, vous pourrez revenir en arrière et disposer d'un bon aperçu.

Definition (Distribution de Bernoulli)

On dit qu'une variable aléatoire X suit une distribution de Bernoulli de paramètre $p \in [0,1]$, noté $X \sim Bern(p)$, si

1. $\mathcal{X} = \{0, 1\}$,
2. $f_X(x) = p1\{x = 1\} + (1 - p)1\{x = 0\}$.

L'espérance, la variance et la fonction génératrice des moments (FGM) de $X \sim Bern(p)$ sont données par

$$\mathbb{E}[X] = p, \qquad \text{var}[X] = p(1-p), \qquad M_X(t) = 1 - p + pe^t.$$

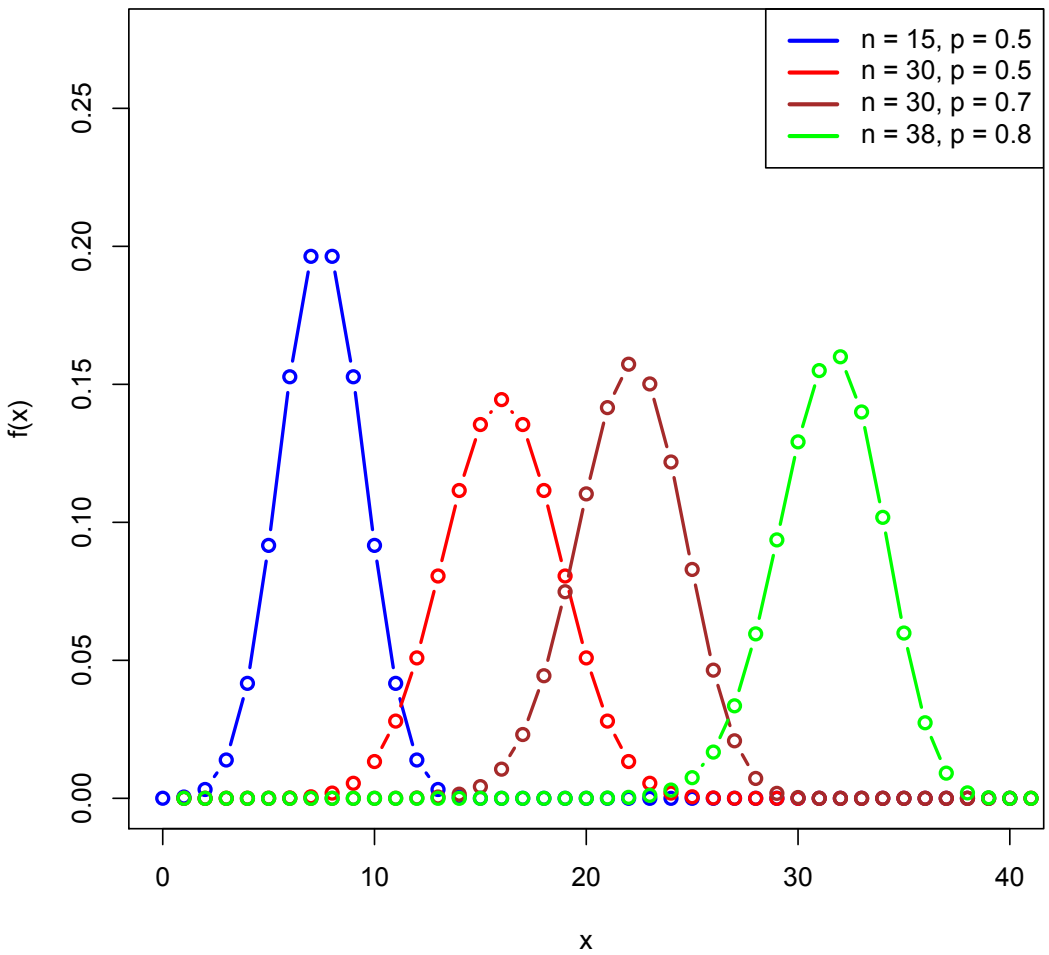
Definition (Distribution binomiale)

On dit qu'une variable aléatoire X suit une distribution binomiale de paramètres $p \in [0,1]$ et $n \in \mathbb{N}$, noté $X \sim Binom(n,p)$, si

1. $\mathcal{X} = \{0, 1, 2, \ldots, n\}$,
2. $f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$.

La moyenne, la variance et la fonction génératrice des moments de $X \sim Binom(n,p)$ sont données par

$$\mathbb{E}[X] = np, \qquad \text{var}[X] = np(1-p), \qquad M_X(t) = (1 - p + pe^t)^n.$$


Binomial Distribution PMF
n = 15, p = 0.5
n = 30, p = 0.5
n = 30, p = 0.7
n = 38, p = 0.8
f(x)
0.00
0.05
0.10
0.15
0.20
0.25
0
10
20
30
40
x

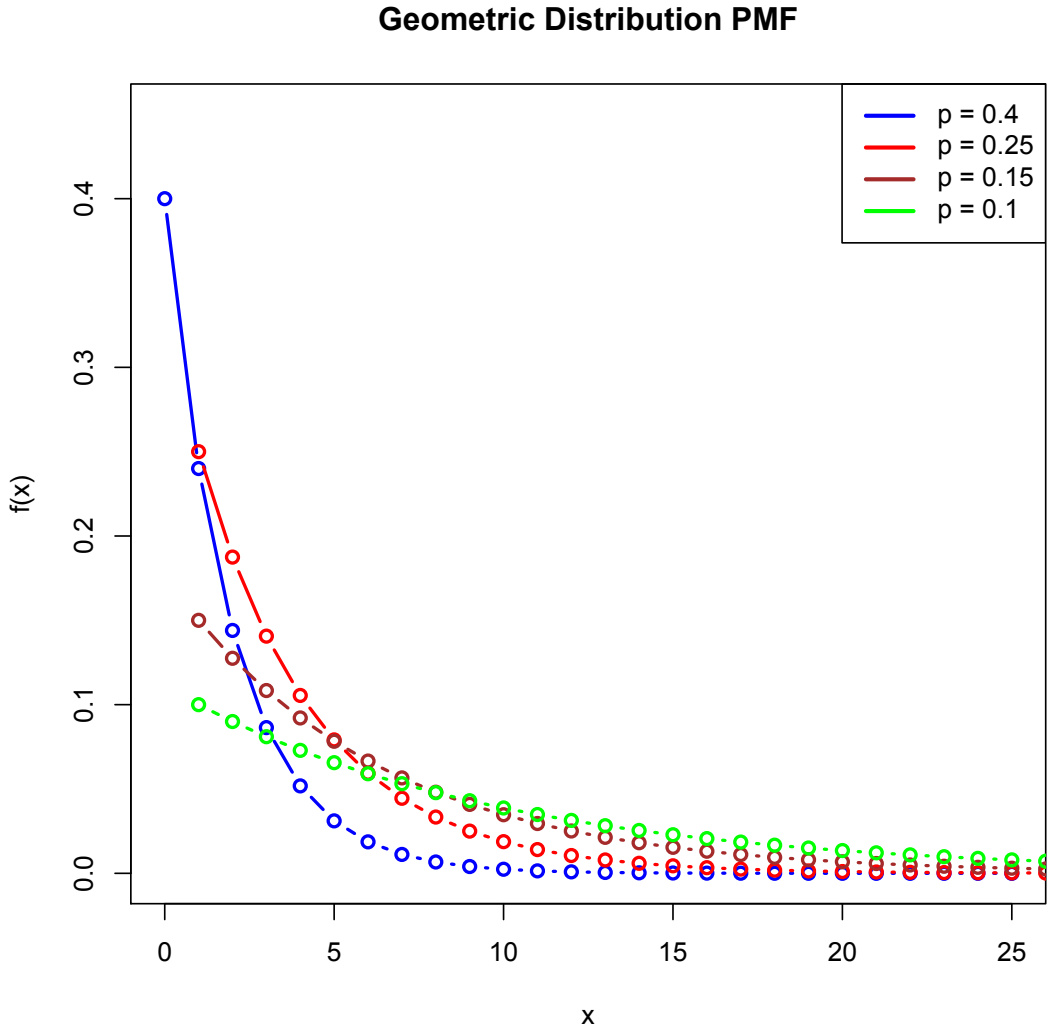
Definition (Distribution géometrique)

Une variable aléatoire X suit une distribution géométrique de paramètre $p \in (0,1]$, noté $X \sim Geom(p)$, si

1. $\mathcal{X} = \{0\} \cup \mathbb{N}$,
2. $f_X(x) = (1-p)^x p$.

La moyenne, la variance et la fonction génératrice des moments de $X \sim Geom(p)$ sont données par

$$\mathbb{E}[X] = \frac{1-p}{p}, \quad \text{var}[X] = \frac{(1-p)}{p^2}, \quad M_X(t) = \frac{p}{1-(1-p)e^t}, \; t < -\log(1-p).$$


Geometric Distribution PMF
p = 0.4
p = 0.25
p = 0.15
p = 0.1
f(x)
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
25
x

Definition (Distribution de Poisson)

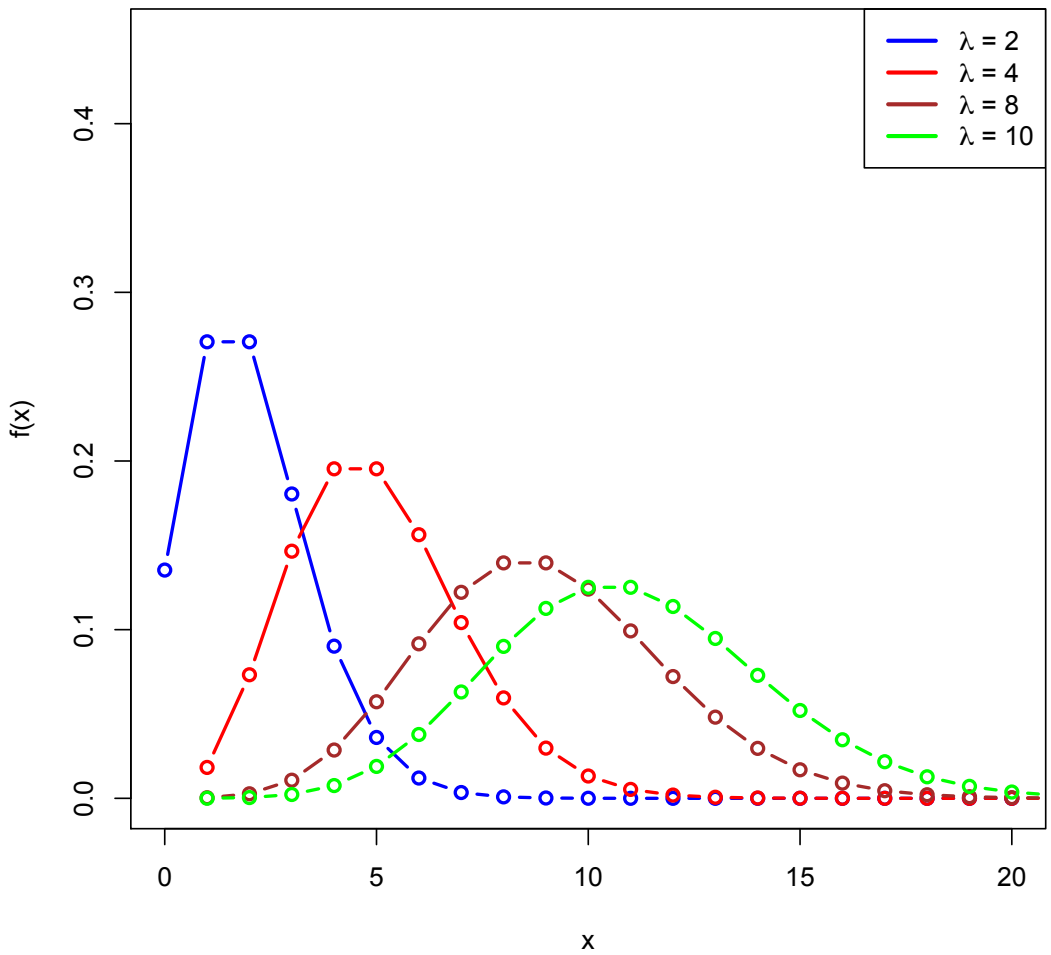
Une variable aléatoire X suit une distribution de Poisson de paramètre $\lambda > 0$, noté $X \sim Poisson(\lambda)$, si

1. $\mathcal{X} = \{0\} \cup \mathbb{N}$,
2. $f_X(x) = e^{-\lambda} \dfrac{\lambda^x}{x!}$.

La moyenne, la variance et la fonction génératrice des moments de $X \sim Poisson(\lambda)$ sont données par

$$\mathbb{E}[X] = \lambda, \qquad \text{var}[X] = \lambda, \qquad M_X(t) = \exp\{\lambda(e^t - 1)\}.$$

Informellement, $Binom(n, p) \to Poisson(\lambda)$ lorsque $n \to \infty$ et $p = \lambda/n$


Poisson Distribution PMF
λ = 2
λ = 4
λ = 8
λ = 10
f(x)
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
x

Definition (Distribution Uniforme)

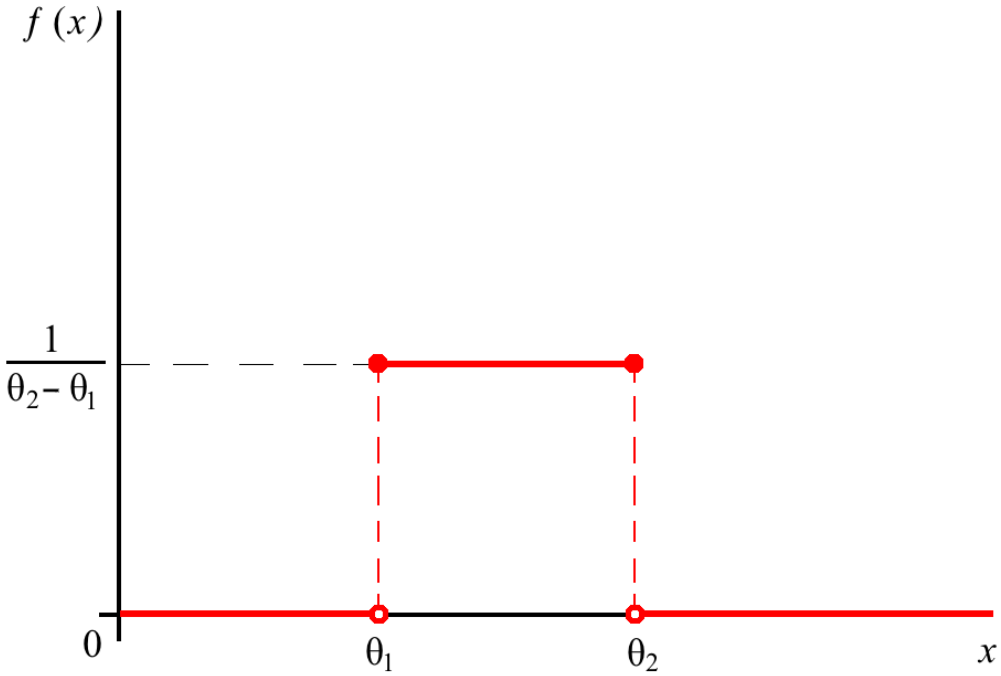
Une variable aléatoire X suit une distribution uniforme de paramètres $-\infty < \theta_1 < \theta_2 < \infty$, noté $X \sim Unif(\theta_1, \theta_2)$, si

$$f_X(x;\theta) = \begin{cases} (\theta_2 - \theta_1)^{-1} & si\ x \in (\theta_1, \theta_2), \\ 0 & sinon. \end{cases}$$

La moyenne, la variance et la fonction génératrice des moments de $X \sim Unif(\theta_1, \theta_2)$ sont données par

$$\mathbb{E}[X] = (\theta_1 + \theta_2)/2, \quad \text{var}[X] = (\theta_2 - \theta_1)^2/12, \quad M_X(t) = \frac{e^{t\theta_2} - e^{t\theta_1}}{t(\theta_2 - \theta_1)},\ t \neq 0, M(0) = 1.$$

Densité uniforme



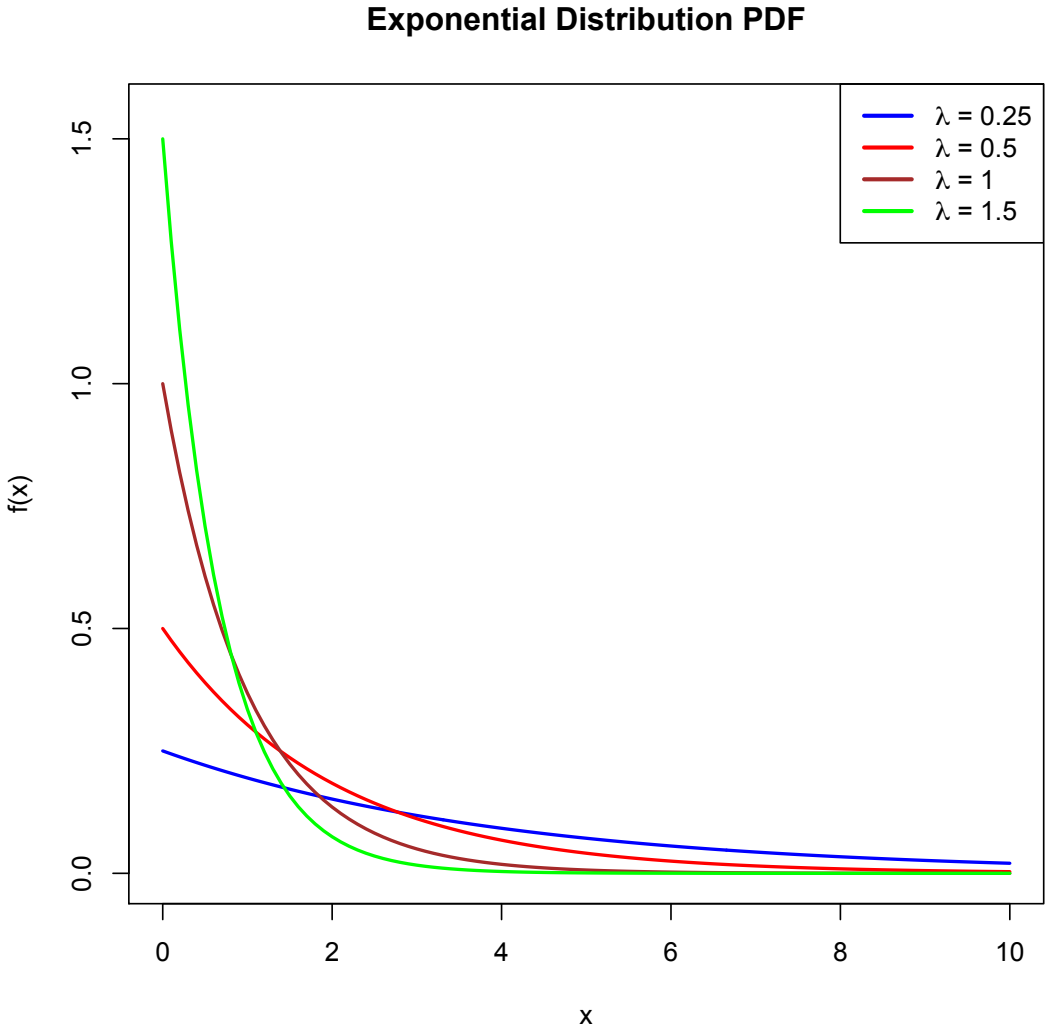
Definition (Distribution exponentielle)

Une variable aléatoire X suit une distribution exponentielle de paramètre $\lambda > 0$, noté $X \sim Exp(\lambda)$, si

$$f_X(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & si\ x \geq 0 \\ 0 & si\ x < 0. \end{cases}$$

La moyenne, la variance et la fonction génératrice des moments $X \sim Exp(\lambda)$ sont données par

$$\mathbb{E}[X] = \lambda^{-1}, \qquad \text{var}[X] = \lambda^{-2}, \qquad M_X(t) = \frac{\lambda}{\lambda - t}, \quad t < \lambda.$$


Exponential Distribution PDF
λ = 0.25
λ = 0.5
λ = 1
λ = 1.5
f(x)
0.0
0.5
1.0
1.5
0
2
4
6
8
10
x

Definition (Distribution normale)

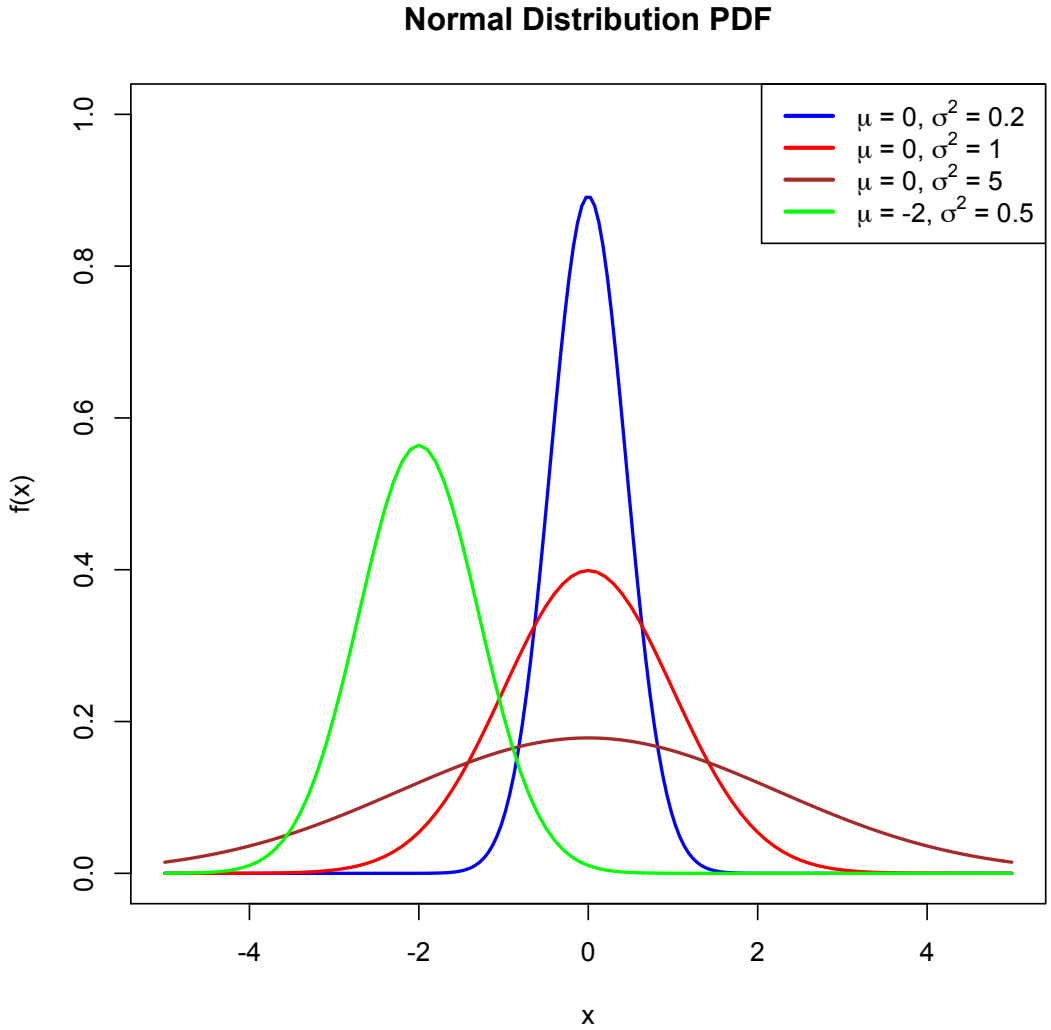
Une variable aléatoire X suit une distribution normale de paramètres $\mu \in \mathbb{R}$ et $\sigma^2 > 0$ (respectivement le paramètre moyenne et le paramètre variance), noté $X \sim N(\mu, \sigma^2)$, si

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad x \in \mathbb{R}.$$

La moyenne, la variance et la fonction génératrice des moments de $X \sim N(\mu, \sigma^2)$ sont données par

$$\mathbb{E}[X] = \mu, \qquad \text{var}[X] = \sigma^2, \qquad M_X(t) = \exp\{t\mu + t^2\sigma^2/2\}.$$

Dans le cas spécial $Z \sim N(0,1)$, nous utilisons la notation $\varphi(z) = f_Z(z)$ et $\Phi(z) = F_Z(z)$, et nous les appelons respectivement la *fonction de densité normale centrée réduite* (ou *fonction de densité normale standard*) et la *fonction de répartition normale centrée réduite* (ou *fonction de répartition normale standard*).


Normal Distribution PDF
μ = 0, σ² = 0.2
μ = 0, σ² = 1
μ = 0, σ² = 5
μ = -2, σ² = 0.5
f(x)
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
x

Transformation des lois de probabilité (ou “propagation de l’incertitude”)

- Souvent: nous avons un modèle pour un phénomène aléatoire X
- Mais nous somme plutôt intéressés par un autre aspect de ce phénomène, disons $g(X)$, où g est une fonction connue.

Example

Supposons que R est une variable aléatoire positive représentant le rayon de couverture d'une antenne Wireless et considérons que $R \sim Unif[a, b]$, pour $0 < a < b$.

Quelle est la distribution de l'aire de couverture $A = \pi R^2$? □

Modèles de probabilité transformés

Comment la distribution d'une variable aléatoire X est transformée, lorsque la variable aléatoire X est transformée?

Lemma

Soit X une variable aléatoire discrète, et $Y = g(X)$. Alors, le support de Y est $\mathcal{Y} = g(\mathcal{X})$ et

$$F_Y(y) = \mathbb{P}[g(X) \leq y] = \sum_{x \in \mathcal{X}} f_X(x) 1\{g(x) \leq y\}, \qquad \forall y \in \mathcal{Y}$$

$$f_Y(y) = \mathbb{P}[g(X) = y] = \sum_{x \in \mathcal{X}} f_X(x) 1\{g(x) = y\}, \qquad \forall y \in \mathcal{Y}.$$

- Preuve = enoncé!
- Cas continu: plus compliqué:
  1. Si g pas monotone: au cas-par-cas.
  2. Si g est monotone: on a des résultats généraux.

Example (La normale standard au carré a une distribution χ^2_1)

Soit $Z \sim N(0,1)$. Nous voulons trouver la distribution de $Y = Z^2$. Notez que $F_Y(y) = \mathbb{P}[Y \leq y] = 0$ si $y < 0$. Pour $y \geq 0$ nous avons :

$$\begin{aligned} F_Y(y) &= \mathbb{P}[Z^2 \leq y] = \mathbb{P}[|Z| \leq \sqrt{y}] = \mathbb{P}[-\sqrt{y} \leq Z \leq \sqrt{y}] \\ &= \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = \Phi(\sqrt{y}) - (1 - \Phi(\sqrt{y})) = 2\Phi(\sqrt{y}) - 1. \end{aligned}$$

Nous pouvons aussi trouver la densité en dérivant :

$$\begin{aligned} f_Y(y) &= 2\frac{d}{dy}\Phi(\sqrt{y}) = 2\frac{d}{d\sqrt{y}}\Phi(\sqrt{y})\frac{d}{dy}\sqrt{y} \\ &= 2\phi(\sqrt{y})\frac{y^{-1/2}}{2} = 2\frac{1}{\sqrt{2\pi}}e^{-y/2}\frac{y^{-1/2}}{2} \\ &= \frac{1}{\sqrt{2}\sqrt{\pi}}e^{-y/2}y^{-1/2} \end{aligned}$$

La dernière expression definie la densité d'une loi appellée "χ^2_1". Alors:

$$\boxed{Z \sim N(0,1) \implies Z^2 \sim \chi^2_1.}$$

Lemma

Soit X une variable aléatoire continue sur $\mathcal{X} \subseteq \mathbb{R}$ et soit $g : \mathcal{X} \to \mathbb{R}$ une

1. *monotone,*
2. *continûment dérivable,*
3. *de derivée jamais nulle.*

Soit $Y = g(X)$. Alors, le support de Y est $\mathcal{Y} = g(\mathcal{X})$ et

- *Si g est croissante, alors*

$$F_Y(y) = F_X(g^{-1}(y)).$$

- *Si g est décroissante, alors*

$$F_Y(y) = 1 - F_X(g^{-1}(y)).$$

Dans les deux cas, nous aurons :

$$f_Y(y) = \left| \frac{\partial}{\partial y} g^{-1}(y) \right| f_X(g^{-1}(y)), \qquad y \in \mathcal{Y}.$$

Proof.

Considérons premièrement le cas où g' est positive partout sur $\mathcal{X}$ (g est croissante). Cela signifie que $x \leq y \iff g(x) \leq g(y)$. Alors, pour $y \in \mathcal{Y}$,

$$F_Y(y) = \mathbb{P}[g(X) \leq y] = \mathbb{P}[X \leq g^{-1}(y)] = F_X(g^{-1}(y)).$$

Ainsi,

$$f_Y(y) = \frac{\partial}{\partial y} F_Y(y) = \frac{\partial}{\partial y} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \frac{\partial}{\partial y} g^{-1}(y) = f_X(g^{-1}(y)) \left| \frac{\partial}{\partial y} g^{-1}(y) \right|,$$

où la dernière égalité vient du fait que g' est positive partout. Considérons maintenant le cas où g est décroissante (et donc g' est négative partout). Cela signifie que $x \leq y \iff g(x) \geq g(y)$. Alors, pour $y \in \mathcal{Y}$,

$$1 - F_Y(y) = \mathbb{P}[g(X) > y] = \mathbb{P}[X < g^{-1}(y)] = F_X(g^{-1}(y)) - \underbrace{\mathbb{P}[X = g^{-1}(y)]}_{=0}.$$

Cependant $f_Y(y) = -\frac{\partial}{\partial y}(1 - F_Y(y))$, ainsi

$$f_Y(y) = -\frac{\partial}{\partial y}(1 - F_Y(y)) = -f_X(g^{-1}(y)) \frac{\partial}{\partial y} g^{-1}(y) = f_X(g^{-1}(y)) \left| \frac{\partial}{\partial y} g^{-1}(y) \right|,$$

puisque $-g'$ est positive partout. Ceci complète la preuve. □

Corollary (Transformations affines)

Soit X une variable aléatoire et $Y = g(X)$. Si $g(x) = ax + b$, $a \neq 0$, alors

$$\forall y \in \mathcal{Y}, \qquad F_Y(y) = \begin{cases} F_X\left(\frac{y-b}{a}\right) & a > 0, \\ 1 - F_X\left(\frac{y-b}{a}\right) + \mathbb{P}\left(X = \frac{y-b}{a}\right) & a < 0, \end{cases}$$

avec $\mathbb{P}\left(X = \frac{y-b}{a}\right) = 0$ si X est une variable aléatoire continue. Ainsi, pour $y \in \mathcal{Y}$:

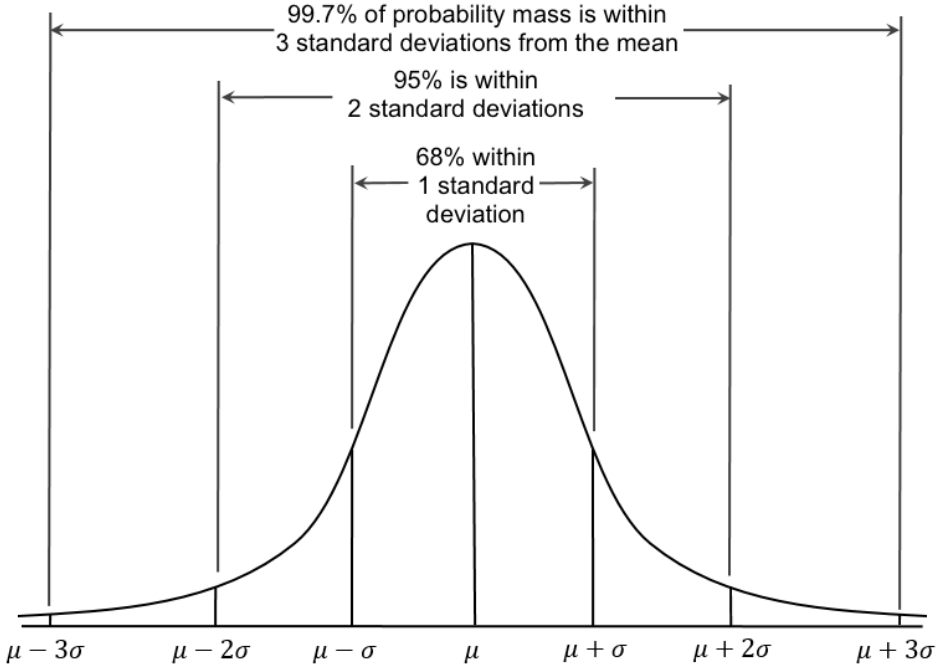
1. $f_Y(y) = |a^{-1}| f_X\left(\dfrac{y-b}{a}\right)$, *si X est continue,*
2. $f_Y(y) = f_X\left(\dfrac{y-b}{a}\right)$, *si X est discrète.*

Example (Transformations affines de la distribution normale)

Soit $X \sim N(\mu, \sigma^2)$, $a \neq 0$. Alors $aX + b \sim N(a\mu + b, a^2\sigma^2)$. Par conséquent, si $X \sim N(\mu, \sigma^2)$, alors

$$F_X(x) = \Phi\left(\frac{x - \mu}{\sigma}\right),$$

où Φ est la fonction de répartition standard,
$\Phi(u) = \int_{-\infty}^{u} (2\pi)^{-1/2} \exp\{-z^2/2\} dz$, qui est, on le rappelle, la fonction de répartition d'une variable aléatoire $Z \sim N(0, 1)$.



Étant donné une probabilité $\alpha \in (0,1)$, quel est le (plus petit) réel x tel que $\mathbb{P}[X \leq x] = \alpha$?

Soit X une variable aléatoire et F_X sa fonction de répartition. Nous définissons la fonction quantile de X (ou équivalemment de F_X) comme la fonction

$$F_X^- : (0,1) \to \mathbb{R}$$

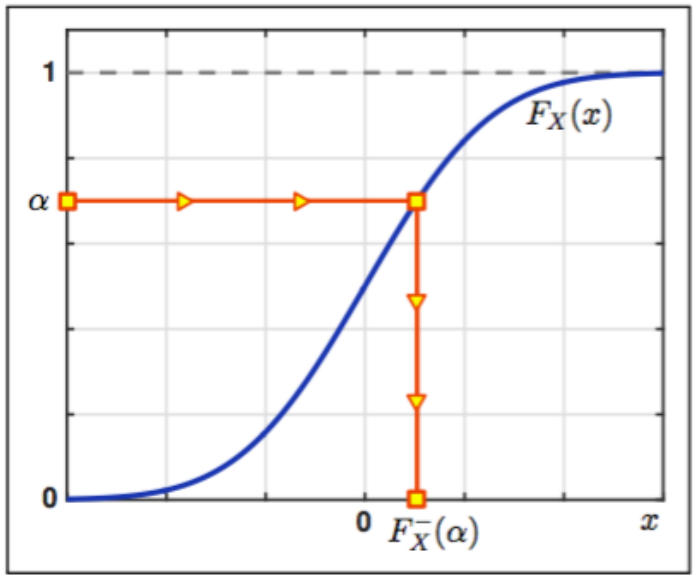
$$F_X^-(\alpha) = \inf\{t \in \mathbb{R} : F_X(t) \geq \alpha\}.$$

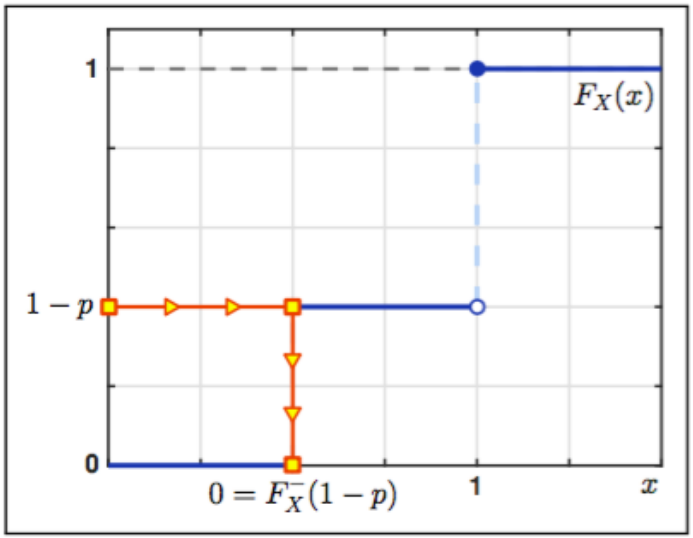
- Si F_X est strictement croissante et continue, alors $F_X^- = F_X^{-1}$.

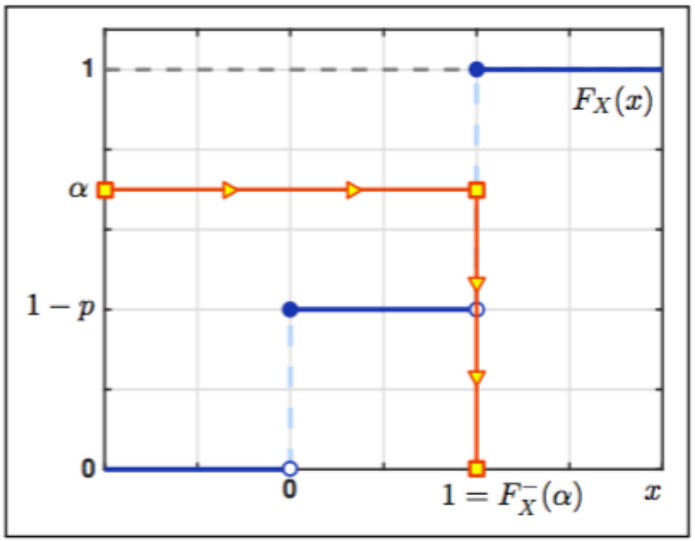
Pour un $\alpha \in (0,1)$ donné, le quantile d'ordre α de X (ou équivalemment de F_X) est le réel

$$q_\alpha = F_X^-(\alpha).$$

Les images qui parlent d'elles-mêmes :







Lemme

Soit $Y \sim \text{Unif}(0,1)$ et F une fonction de répartition. Alors, la fonction de répartition de la variable aléatoire $X = F^{-}(Y)$ est précisément F.

- Exercice : vérifiez la véracité de l'énonce
- Permet de générer des variables ayantes n'importe quelle loi
  - *À condition de pouvoir générer des réalisations de la loi uniforme sur* $[0,1]$.
  - *On peut le faire à l'aide des développements binaires et de tirages de Bernoulli.*
  - *Réduit le problème à un lancer de pièce infini.*

Converse partielle

Soit X une variable aléatoire de fonction de répartition F_X strictement croissante et continue. Alors, $F_X(X) \sim \text{Unif}(0,1)$.

Moments et leurs fonctions génératrices

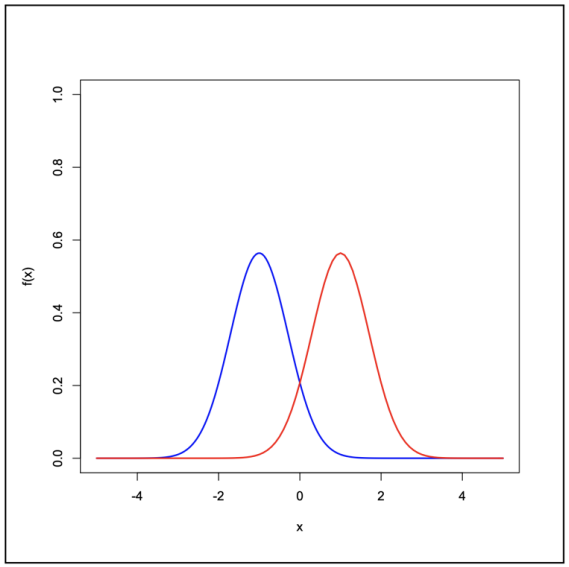
Existe-t-il des résumés numériques qui nous aident à apprécier certaines caractéristiques importantes d'une loi de probabilité — en particulier des aspects clés de sa "forme" ?

On peut voir une loi de probabilité comme la façon dont une masse (de taille totale 1) est distribuée dans l'espace. Et comme pour toute masse, on peut regarder où elle se situe, on peut regarder si elle est serrée ou étalée, on peut regarder ses symétries, et on peut regarder si elle s'étend loin.

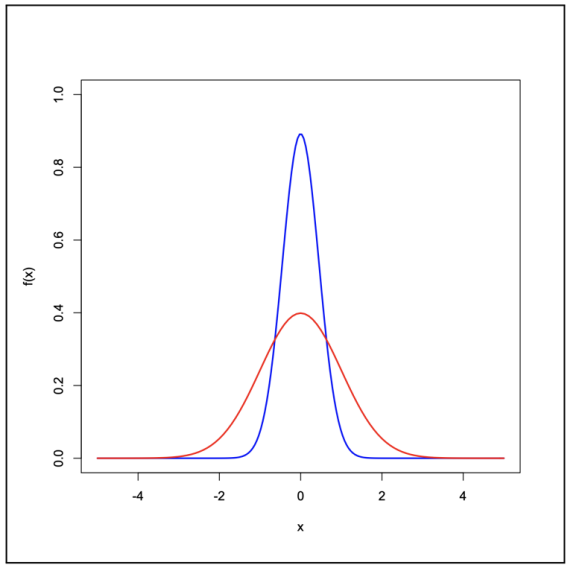
En resumé on veut décrire:

1. Position.
2. Dispersion.
3. Symétrie/Asymétrie.
4. Comportement des Queues.

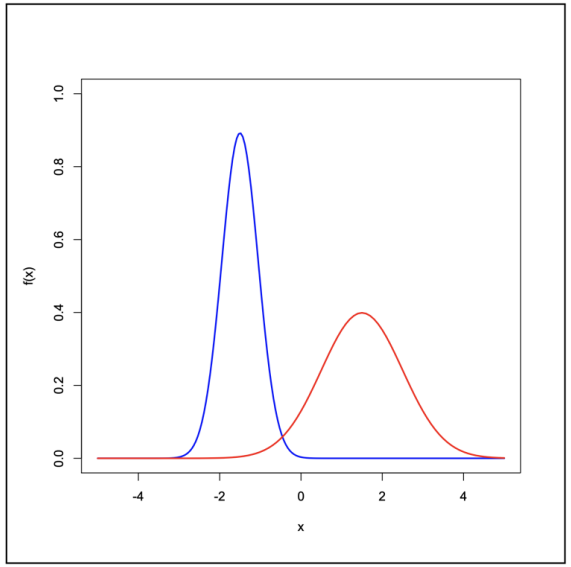
Les moments donnent un moyen mathématique pour capturer ces descripteurs.



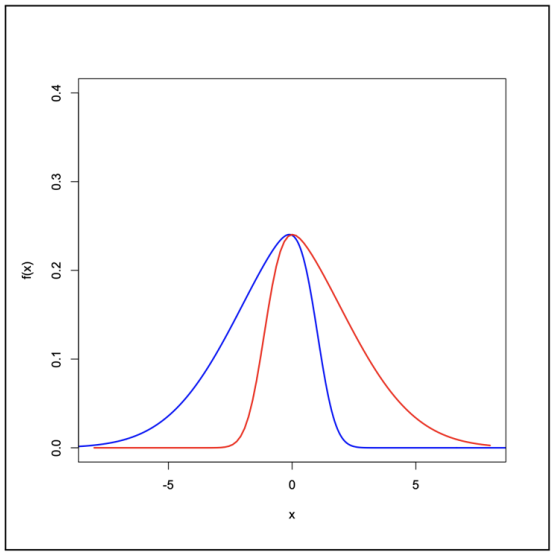
(a) Deux densités de positions différentes.



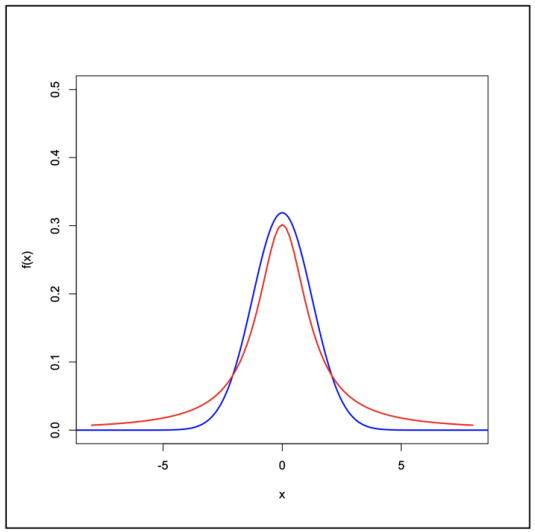
(b) Deux densités de dispersions différentes.



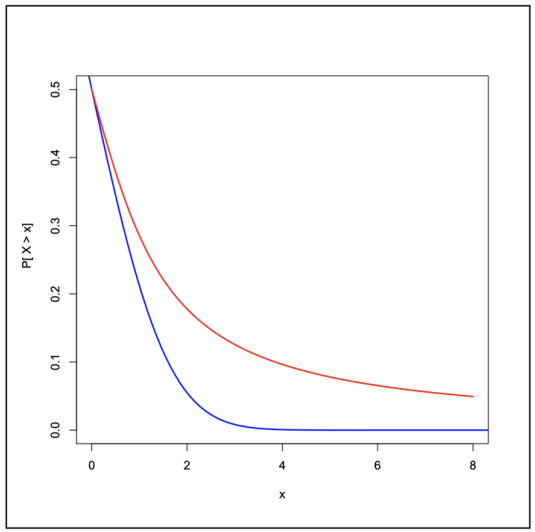
(c) Deux densités qui diffèrent par leur position et leur dispersion.



(d) Deux densités asymétriques : une avec une asymétrie positive (rouge), et une avec une asymétrie négative (bleu).



(e) Une densité. à queue lourde (rouge) et une densité à queue légère (bleu).



(f) Graphique de la fonction $x \mapsto \int_x^\infty f(y)dy$ pour les deux densités de gauche.

L'espérance (ou valeur attendue) d'une variable aléatoire X formalise la notion de "valeur moyenne" prise par cette variable.

- C'est la "moyenne pondérée" de X, avec coefficients $f_X(x)$, $x \in \mathcal{X}$.
- Représente le "centre de gravité" d'une loi, vue comme une masse

Pour les variables continues, elle est définie comme :

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x\, f_X(x)\, dx,$$

à condition que $\int_{\mathbb{R}} |x| f_X(x)\, dx < \infty$.

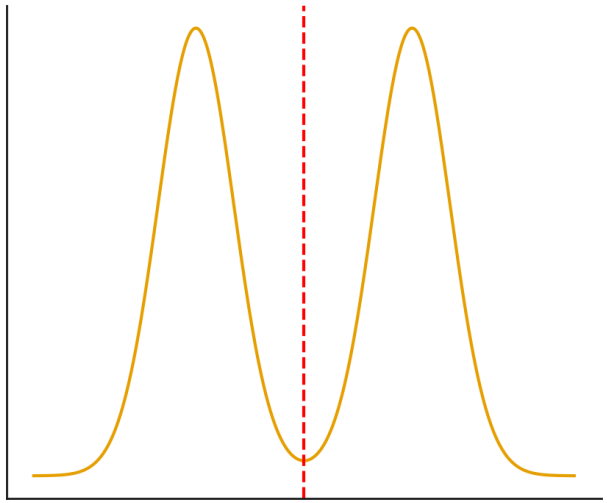
Pour les variables discrètes, elle est définie comme :

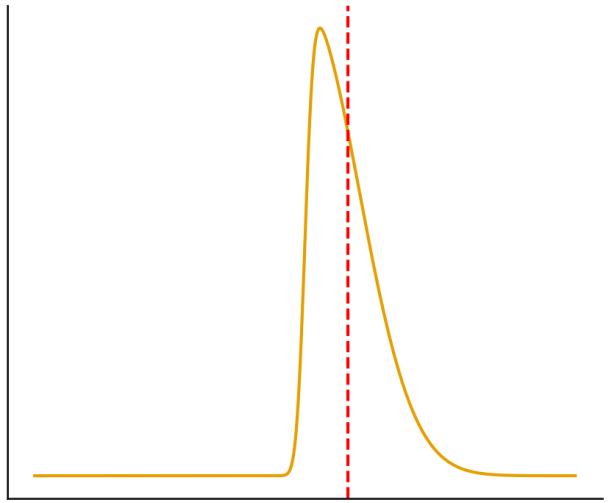
$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x\, f_X(x), \qquad \mathcal{X} = D_{F_X} = \{x \in \mathbb{R} : f_X(x) > 0\}.$$

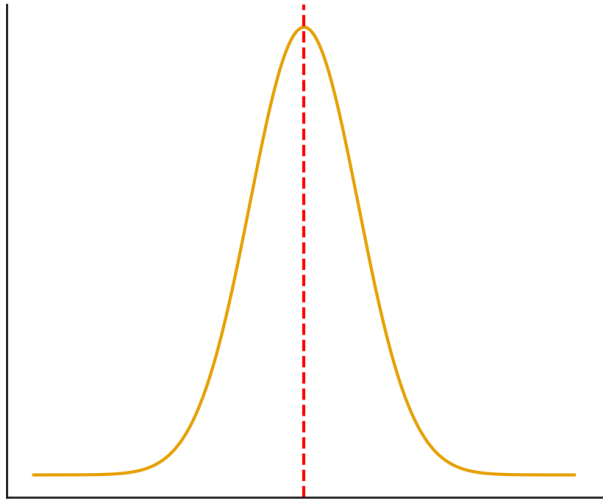
à condition que $\sum_{x \in \mathcal{X}} |x| f_X(x) < \infty$.

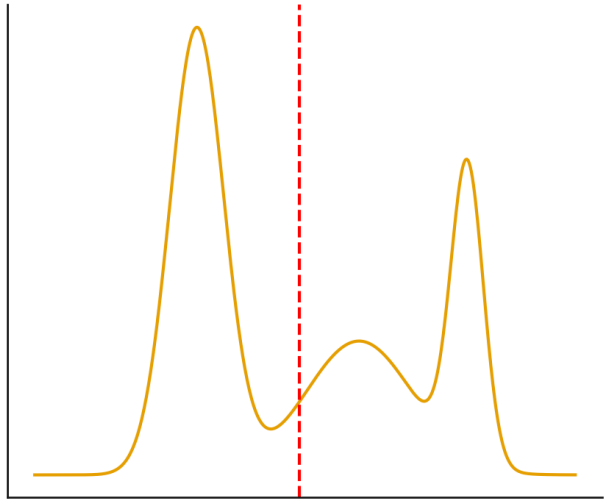
Par la définition, il est immédiat que $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$, $\forall\ a, b \in \mathbb{R}$.

Densités avec la même moyenne (espérance)









Example

Pour les lois que nous avons vu :

- $X \sim \text{Bernoulli}(p) \implies \mathbb{E}[X] = p$
- $X \sim \text{Binom}(n, p) \implies \mathbb{E}[X] = np$
- $X \sim \text{Geom}(p) \implies \mathbb{E}[X] = \frac{1-p}{p}$
- $X \sim \text{Poisson}(\lambda) \implies \mathbb{E}[X] = \lambda$
- $X \sim \text{Unif}(a, b) \implies \mathbb{E}[X] = (a + b)/2$
- $X \sim \text{exp}(\lambda) \implies \mathbb{E}[X] = \frac{1}{\lambda}$
- $X \sim \text{N}(\mu, \sigma^2) \implies \mathbb{E}[X] = \mu$

Faisons explicitement les cas mis en évidence.

Soit $X \sim \text{Bernoulli}(p)$. Alors,

$$\mathbb{E}[X] = \sum_{x \in \{0,1\}} x \, \mathbb{P}(X = x) = 0 \cdot (1 - p) + 1 \cdot p = p.$$

Soit $X \sim \text{Poisson}(\lambda)$, alors $\mathbb{E}[X] = \sum_{k=0}^{\infty} k\,\mathbb{P}(X=k) = \sum_{k=1}^{\infty} k\,e^{-\lambda}\frac{\lambda^k}{k!}$. On réécrit $k/k! = 1/(k-1)!$ et on factorise λ :

$$\mathbb{E}[X] = e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!} = e^{-\lambda}\,\lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}.$$

Posons $j = k-1$: $\mathbb{E}[X] = e^{-\lambda}\,\lambda\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} = e^{-\lambda}\,\lambda\,e^{\lambda} = \lambda$.

Soit $X \sim \text{N}(\mu, \sigma^2)$, et alors $X = \sigma Z + \mu$, $Z \sim \text{N}(0,1)$ dont le densité est

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

Alors

$$\mathbb{E}[Z] = \int_{-\infty}^{\infty} z\,\phi(z)\,dz.$$

L'intégrande $z\,\phi(z)$ est *impaire* (car ϕ est paire), donc l'intégrale vaut 0 : $\mathbb{E}[Z] = 0$. Maintenant si $X = \sigma Z + \mu$ $\mathbb{E}[X] = \mathbb{E}[\sigma Z + \mu] = \sigma\,\mathbb{E}[Z] + \mu = \mu$.

Si l'espérance $\mathbb{E}[X]$ de X donne le centre de gravité de sa loi, qu'en est-il de :

- l'espérance $\mathbb{E}[X^2]$ de X^2 ?
- l'espérance $\mathbb{E}[X^3]$ de X^3 ?

$\vdots$

- l'espérance $\mathbb{E}[X^k]$ de X^k ?
- plus généralement, l'espérance $\mathbb{E}[g(X)]$ de $g(X)$?

Commençons par énoncer un théorème très utile : il n'est pas nécessaire de déterminer la loi de la variable transformée $Y = g(X)$ puis de calculer $\mathbb{E}[Y]$;

Théorème (Espérance d'une transformation)

Pour $g : \mathcal{X} \to \mathbb{R}$ une transformation sur le support $\mathcal{X}$ de X, on a :

Cas discret : $\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) f_X(x)$, à condition que $\sum_{x \in \mathcal{X}} |g(x)| f_X(x) < \infty$.

Cas continu : $\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x)\, dx$, à condition que $\int_{\mathbb{R}} |g(x)| f_X(x)\, dx < \infty$.

La variance d'une variable aléatoire X est définit comme

$$\text{var}(X) = \mathbb{E}\{[X - \mathbb{E}(X)]^2\} = \cdots = \mathbb{E}(X^2) - \mathbb{E}(X)^2,$$

à condition que $\mathbb{E}(X^2) < \infty$.

- C'est l'erreur quadratique moyenne, lorsqu'on prévoit X par $\mathbb{E}[X]$.
- Représente le moment d'inertie relatif au centre de masse.
- $\text{var}(X) \geq 0$
- $\text{var}(X) = 0 \implies X$ est constante; en fait que $\mathbb{P}[X = \mathbb{E}(X)] = 1$.
- la l'écart type de X est défini comme $\sqrt{\text{var}(X)} \geq 0$.
- $\text{var}(aX + b) = a^2\text{var}(X)$ pour tout $a, b \in \mathbb{R}$. constantes

Example

- Si $X \sim \text{Poiss}(\lambda)$, alors $\text{var}(X) = \lambda$.
- Si $X \sim \text{Binom}(n, p)$, alors $\text{var}(X) = np(1-p)$.
- Si $X \sim \text{N}(\mu, \sigma^2)$, alors $\text{var}(X) = \sigma^2$.

La standardisation d'une variable aléatoire X consiste à construire la variable

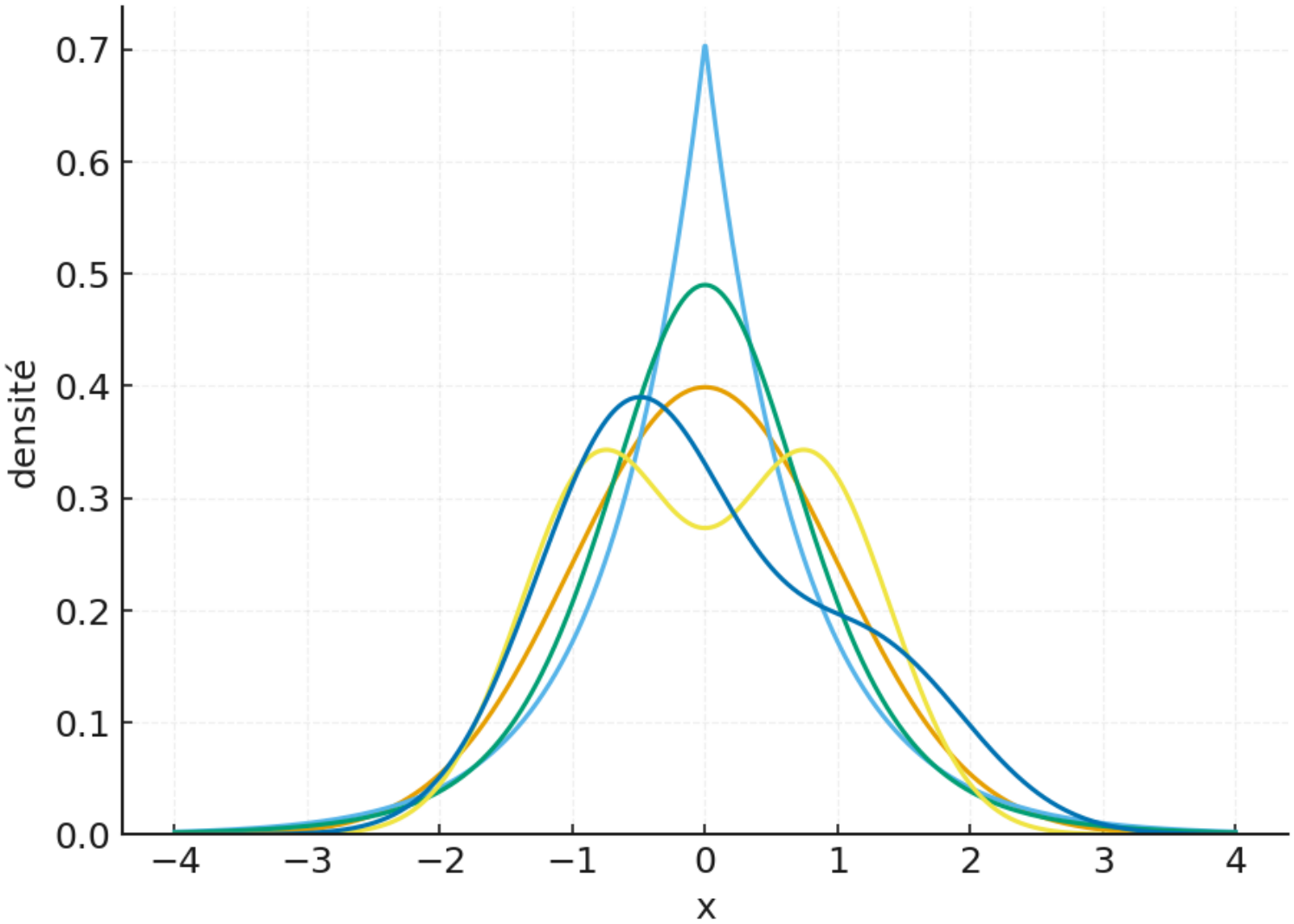
$$Z = \frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{var}(X)}},$$

à condition que $0 < \mathrm{var}(X) < \infty$.

- Z est une version centrée-réduite de X.
- Par construction, $\mathbb{E}[Z] = 0$ et $\mathrm{var}(Z) = 1$.
- Permet de comparer des variables de natures ou d'unités différentes.
- Dans le cas $X \sim \mathrm{N}(\mu, \sigma^2)$, on obtient $Z \sim \mathrm{N}(0, 1)$

(ce qui nous permet de calculer de probabilités pour un loi normale quelconque)

Densités standardisées (avec la même moyenne 0 et la même variance 1)



Le coefficient d'asymétrie d'une variable aléatoire X est définie comme

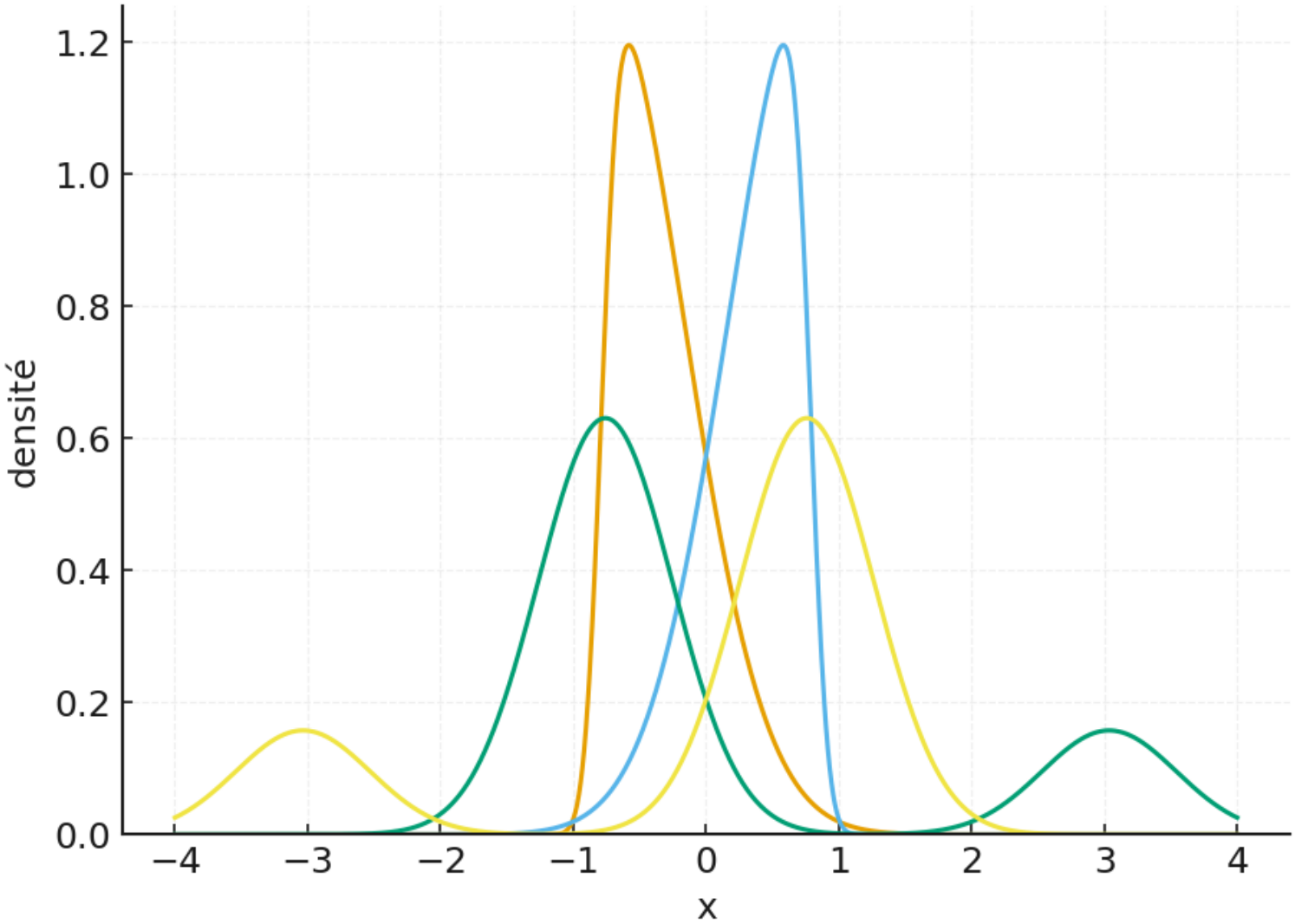
$$\mathbb{E}\left[\left(\frac{X-\mathbb{E}[X]}{\sqrt{\mathrm{var}(X)}}\right)^3\right], \quad \text{à condition que } \mathbb{E}[|X|^3] < \infty.$$

- C'est le moment d'ordre 3 de la variable standardisée.
- Mesure l'asymétrie de la loi par rapport à sa moyenne.
- $= 0$ signifie que la loi est symétrique autour de la moyenne.
- > 0 : queue droite plus lourde (asymétrie à droite).
- < 0 : queue gauche plus lourde (asymétrie à gauche).

Example

- Si $X \sim \mathrm{N}(\mu, \sigma^2)$, alors le coefficient vaut 0.
- Si $X \sim \mathrm{Exp}(\lambda)$, alors le coefficient vaut 2 (asymétrie à droite).

Densités avec la même moyenne, la même variance et la même valeur absolue de l'asymétrie (skewness)



Deux questions :

- Comme plus l'ordre du moment est élevé, plus la caractérisation est fine : le fait de connaître tous les moments détermine-t-il de manière unique la loi de probabilité sous-jacente ?

- Existe-t-il un moyen plus simple de calculer le k-ième moment qu'une intégration directe ? (et donc, d'approximer l'espérance d'une fonction suffisamment régulière $g(X)$ par un développement de Taylor ?)

Il se trouve que la réponse à ces deux questions est *pas tout à fait toujours*.

Mais sous des conditions fortes sur les queues, la réponse est un grand oui — et elle est donnée par la fonction génératrice des moments.

La condition de queue est appelée :

Condition d'intégrabilité exponentielle

$$\exists\, \epsilon > 0 : \ \mathbb{E}\left[e^{tX}\right] < \infty, \quad \text{pour } t \in (-\epsilon, \epsilon).$$

La fonction génératrice des moments (FGM) $M_X(t) : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ d'une variable aléatoire scalaire X est définie par

$$\boxed{M_X(t) = \mathbb{E}\left[e^{tX}\right], \ t \in \mathbb{R}.}$$

Elle n'a pas besoin d'être finie pour $t \neq 0$. Mais lorsqu'elle est finie dans un voisinage de zéro (intégrabilité exponentielle), de belles propriétés apparaissent :

Theorem

Soient X et Y deux variables aléatoires scalaires, et supposons que $M_X(t) < \infty$ et $M_Y(t) < \infty$ pour tout $t \in I = (-\epsilon, \epsilon)$, pour un certain $\epsilon > 0$. Alors :

1. *M_X est indéfiniment dérivable sur I.*
2. *$\mathbb{E}[|X|^k] < \infty$ et $\mathbb{E}[X^k] = \frac{d^k M_X}{dt^k}(0)$, pour tout $k \geq 1$.*
3. *$F_X = F_Y$ sur $\mathbb{R} \iff M_X = M_Y$ sur I.*

Nous n'allons pas prouver le théorème, mais donner quelques intuitions pour comprendre pourquoi il a du sens.

Intuition à retenir : l'intégrabilité exponentielle fournit un contrôle uniforme qui autorise (via des résultats d'analyse) l'échange "espérance $\leftrightarrow$ dérivation".

Ainsi, si l'on peut passer la dérivée *à l'intérieur* de l'espérance :

$$\frac{d}{dt}\, e^{tX} = Xe^{tX}, \quad \frac{d^k}{dt^k}\, e^{tX} = X^k e^{tX},$$

alors, évidemment, M_X est indéfiniment dérivable sur I. De plus, pour $|t| < \varepsilon$,

$$|X^k e^{tX}| \leq C_k\, e^{c|X|} \quad \text{pour des constantes } C_k, c > 0,$$

d'où on obtient (2), et

$$M_X^{(k)}(t) = \mathbb{E}[X^k e^{tX}] \;\Rightarrow\; M_X^{(k)}(0) = \mathbb{E}[X^k].$$

Pour voir pourquoi (3) peut avoir du sens, considérons le cas $\mathcal{X} = \mathcal{Y} = \{0, 1, \ldots, n\}$, avec $n \geq 1$, et posons $p_k = \mathbb{P}[X = k]$, $q_k = \mathbb{P}[Y = k]$. Alors :

$$M_X(t) - M_Y(t) = \sum_{k=0}^{n} (p_k - q_k) e^{tk} = \sum_{k=0}^{n} (p_k - q_k) \left(e^t\right)^k = \sum_{k=0}^{n} (p_k - q_k) z^k = P(z).$$

Lorsque $t \in I$, on a $e^t \in (e^{-\epsilon}, e^{\epsilon})$. Ainsi $M_X(t) - M_Y(t)$ s'annule sur I si et seulement si le polynôme $P(z)$ s'annule sur un intervalle ouvert, ce qui n'arrive que lorsque tous ses coefficients sont nuls, c'est à dire $p_k = q_k$, $k \in \{0, \ldots, n\}$.

Example

Soit $X \sim \mathrm{Geom}(p)$, alors

$$M_X(t) = \mathbb{E}[e^{tX}] = \sum_{k=0}^{\infty} e^{tk}(1-p)^k p.$$

On reconnaît une série géométrique de raison $r = (1-p)e^t$, qui converge si $|(1-p)e^t| < 1$, c.-à-d. $t < -\ln(1-p)$. Ainsi,

$$M_X(t) = \frac{p}{1-(1-p)e^t}, \qquad t < -\ln(1-p).$$

Soit $X \sim \mathrm{Exp}(\lambda)$, alors

$$M_X(t) = \mathbb{E}[e^{tX}] = \int_0^{\infty} e^{tx}\, \lambda e^{-\lambda x}\, dx = \lambda \int_0^{\infty} e^{-(\lambda - t)x}\, dx.$$

Cette intégrale converge ssi $t < \lambda$, et vaut

$$M_X(t) = \frac{\lambda}{\lambda - t}, \qquad t < \lambda.$$

Soit $X \sim \text{Poisson}(\lambda)$. Alors sa fonction génératrice des moments est

$$M_X(t) \;=\; \mathbb{E}[e^{tX}] = \sum_{k=0}^{\infty} e^{tk}\, e^{-\lambda}\, \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} = \exp\big(\lambda(e^t - 1)\big).$$

où on a reconnu la série exponentielle. Et elle est bien définie pour tout $t \in \mathbb{R}$.
Maintenant, considérons $X \sim \text{N}(\mu, \sigma^2)$. Écrivons $X = \mu + \sigma Z$ avec $Z \sim \text{N}(0,1)$:

$$M_X(t) = \mathbb{E}[e^{t(\mu+\sigma Z)}] = e^{\mu t}\, \mathbb{E}[e^{t\sigma Z}].$$

Or, pour $Z \sim \text{N}(0,1)$,

$$\mathbb{E}[e^{sZ}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sz}\, e^{-z^2/2}\, dz = e^{\,s^2/2}$$

par la complétion du carré $sz - \frac{1}{2}z^2 = -\frac{1}{2}\Big(z^2 - 2sz\Big) = -\frac{1}{2}\Big[(z-s)^2 - s^2\Big]$. En prenant $s = t\sigma$, on obtient

$$M_X(t) = e^{\mu t}\, e^{(t\sigma)^2/2} = \exp\Big(\mu t + \tfrac{1}{2}\sigma^2 t^2\Big).$$

D'abord, notons (exercice) que pour toute variable aléatoire,

$$\mathbb{E}[|X|] = \int_0^\infty \mathbb{P}[|X| > t]\,dt.$$

Soit Y une variable aléatoire positive. Alors, pour tout $\epsilon > 0$,

$$\mathbb{P}[Y \geq \epsilon] = \frac{1}{\epsilon}\int_0^\epsilon \mathbb{P}(Y \geq \epsilon)\,dt \leq \frac{1}{\epsilon}\int_0^\epsilon \mathbb{P}(Y > t)\,dt \leq \frac{1}{\epsilon}\int_0^\infty \mathbb{P}(Y > t)\,dt = \frac{\mathbb{E}[Y]}{\epsilon} \quad \text{[Markov]}$$

Soit X une variable aléatoire telle que $\mathbb{E}[|X|^k] < \infty$. Alors, pour tout $\epsilon > 0$,

$$\mathbb{P}\Big[|X - \mathbb{E}[X]|^k \geq \epsilon^k\Big] \leq \frac{\mathbb{E}[|X - \mathbb{E}(X)|^k]}{\epsilon^k} \qquad \text{[Tchebychev]}$$

Remarque : Des bornes encore plus fines découlent de la FGM [Chernoff]

Pour toute fonction convexe[2] $\varphi : \mathbb{R} \to \mathbb{R}$, si $\mathbb{E}|\varphi(X)| + \mathbb{E}|X| < \infty$, alors

$$\varphi\Big(\mathbb{E}[X]\Big) \leq \mathbb{E}[\varphi(X)] \qquad \text{[Jensen[3]]}$$

[2] φ est convexe si $\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y)$ pour tout x, y, et $\lambda \in [0,1]$.

[3] Aide-memoire : rappelons que $0 \leq \mathtt{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ et $\varphi(x) = x^2$ est convexe

Entropie et Familles exponentielles

- Comme nous l'avons vu, les moments peuvent décrire des caractéristiques clés de la forme d'une distribution. Mais un nombre fini de moments ne détermine pas entièrement une distribution.
- Si nous choisissons une forme de loi à partir d'un nombre fini de conditions sur les moments, mais que nous ne sommes pas sûrs de sa structure plus fine, alors de nombreux choix restent possibles.
- Existe-t-il un choix canonique ? Peut-être la distribution "la plus désordonné" ou le moins "contraint" ?
  (afin de ne pas introduire de déterminisme arbitraire)

L'entropie d'une variable aléatoire X est définie par

$$H(X) = -\mathbb{E}\Big\{\log f_X(X)\Big\} = \begin{cases} -\displaystyle\sum_{x\in\mathcal{X}} f_X(x)\log\{f_X(x)\}, & \text{si } X \text{ est discrète,} \\ -\displaystyle\int_{-\infty}^{+\infty} f_X(x)\log\{f_X(x)\}\,dx, & \text{si } X \text{ est continue.} \end{cases}$$

Une mesure du désordre intrinsèque ou de l'imprévisibilité d'un système aléatoire.

Considérons le problème variationnel suivant (traitons le cas continu, mais le cas discret est pareil) : déterminer la loi de probabilité f supportée sur $\mathcal{X}$ ayant une entropie maximale

$$H(f) = -\int_{\mathcal{X}} f(x) \log f(x) \, dx$$

sous les contraintes linéaires ("moments généralisées")

$$\int_{\mathcal{X}} T_i(x) f(x) \, dx = \alpha_i, \qquad i = 1, ..., k.$$

Proposition.

Lorsqu'une solution au problème existe, elle est unique et de la forme

$$f(x) = Q(\lambda_1, ..., \lambda_k) \exp\left\{ \sum_{i=1}^{k} \lambda_i \, T_i(x) \right\}.$$

Preuve.

Soit $g(\cdot)$ une densité satisfaisant aussi les contraintes et $X \sim g(\cdot)$. Alors,

$$
\begin{aligned}
H(g) &= -\int_{\mathcal{X}} g(x)\log g(x)\,dx = -\int_{\mathcal{X}} g(x)\log\left[\frac{g(x)}{f(x)}f(x)\right]dx \\
&= \mathbb{E}\{\log[f(X)/g(X)]\} - \int_{\mathcal{X}} g(x)\log f(x)\,dx \\
&\leq \log \mathbb{E}[f(X)/g(X)] - \int_{\mathcal{X}} g(x)\log f(x)\,dx \quad \text{[Jensen]} \\
&= \underbrace{\log \int_{\mathcal{X}} f(x)\,dx}_{=0} - \log Q \underbrace{\int_{\mathcal{X}} g(x)\,dx}_{=1} - \int_{\mathcal{X}} g(x)\left(\sum_{i=1}^{k}\lambda_i T_i(x)\right)dx.
\end{aligned}
$$

Mais g satisfait aussi les contraintes de moments, donc le dernier terme est

$$
= -\log Q - \int_{\mathcal{X}} f(x)\left(\sum_{i=1}^{k}\lambda_i T_i(x)\right)dx = -\int_{\mathcal{X}} f(x)\log f(x)\,dx = H(f).
$$

Pour l'unicité de la solution, on observe que l'inégalité devient égalité lorque $\mathbb{E}\{\log[f(X)/g(X)]\} = 0$.

Montrons que cela arrive ssi $f(x) = g(x)$ sur $\mathcal{X}$. On va utiliser l'inégalité $\log u \leq u - 1$ (avec égalité si et seulement si $u = 1$).

En posant $u(x) = f(x)/g(x)$ (sur $\mathcal{X}$), on obtient

$$-\log \frac{f(x)}{g(x)} \geq 1 - \frac{f(x)}{g(x)}, \quad x \in \mathcal{X}.$$

En multipliant par $g(x)$ et en intégrant,

$$\int_{\mathcal{X}} g(x) \left(-\log \frac{f(x)}{g(x)} \right) dx \geq \int (g(x) - f(x))\, dx = 1 - 1 = 0.$$

L'égalité $\log u = u - 1$ a lieu si et seulement si $u(X) = 1$, partout sur $\mathcal{X}$, ce qui signifie $\frac{f(x)}{g(x)} = 1$ sur le support commun $\mathcal{X}$. □

Une loi de probabilité appartient à une famille exponentielle à k paramètres, si sa densité (ou fonction de masse) admet la représentation

$$f(y) = \exp\left\{\sum_{i=1}^{k} \phi_i \, T_i(y) - \gamma(\phi_1, ..., \phi_k) + S(y)\right\}$$

où :

1. $\phi = (\phi_1, ..., \phi_k)$ est un paramètre k-dimensionnel dans $\Phi \subseteq \mathbb{R}^k$;
2. $T_i : \mathcal{Y} \to \mathbb{R}$, $i = 1, ..., k$, $S : \mathcal{Y} \to \mathbb{R}$, et $\gamma : \mathbb{R}^k \to \mathbb{R}$, sont à valeurs réelles;
3. Le support $\mathcal{Y}$ de f ne dépend pas de ϕ.

Classe très riche de modèles (parfois avec certains paramètres fixés pour satisfaire la dernière condition) : Binomiale, Poisson, géometrique, exponentielle, Gamma, Gaussienne, Pareto, Weibull, Laplace, log-Normale, gaussienne inverse, gamma inverse, normale-gamma, Beta, Multinomiale...

La loi de Boltzmann est un cas particulier de la famille exponentielle, centrale en physique, où $k = 1$, $T_1(y) = -E(y)$ est le negatif de l'énergie, $\varphi = \beta$ est l'inverse de la température, et $S(y) = 1$.

- $\boldsymbol{\phi} = (\phi_1, ..., \phi_k)^\top$ est appelé le paramètre naturel.
- Mais en changeant de paramètre, on peut écrire la famille exponentielle sous d'autres formes.
- "Naturel" au sens mathématique – le paramètre usuel $\boldsymbol{\theta} = \eta^{-1}(\boldsymbol{\phi})$ est souvent différent.

Paramétrisation naturelle vs usuelle

$$\exp\left\{\sum_{i=1}^{k} \phi_i\, T_i(y) - \gamma(\boldsymbol{\phi}) + S(y)\right\} = \exp\left\{\sum_{i=1}^{k} \eta_i(\boldsymbol{\theta})\, T_i(y) - d(\boldsymbol{\theta}) + S(y)\right\}.$$

où $\eta : \mathbb{R}^k \to \mathbb{R}^k$ est une application C^2 telle que

$$\boldsymbol{\phi} = \eta(\boldsymbol{\theta})$$

et donc $\gamma(\boldsymbol{\phi}) = \gamma(\eta(\boldsymbol{\theta})) = d(\boldsymbol{\theta})$, avec $d = \gamma \circ \eta$.

- Paramétrisation naturelle : excellente pour la manipulation mathématique.
- Paramétrisation usuelle : plus intuitive dans le contexte des applications.

Example (Famille exponentielle binomiale)

Soit $Y \sim \text{Binom}(n, p)$. On observe que :

$$\binom{n}{y} p^y (1-p)^{n-y} = \exp\left\{ \log\left(\frac{p}{1-p}\right) y + n \log(1-p) + \log\binom{n}{y} \right\}.$$

On définit

$$\phi = \log\left(\frac{p}{1-p}\right), \quad T(y) = y,$$

$$S(y) = \log\binom{n}{y}, \quad \gamma(\phi) = n\log(1+e^{\phi}) = -n\log(1-p).$$

En gardant n fixé et en laissant seulement p varier, le support de f ne dépend pas de ϕ et on obtient une famille exponentielle à 1 paramètre. Remarquons que :

$$p = \frac{e^{\phi}}{1+e^{\phi}} \qquad \& \qquad \phi = \underbrace{\log\left(\frac{p}{1-p}\right)}_{=\eta(p)}.$$

Donc le paramètre usuel est $p \in (0,1)$, mais le paramètre naturel est $\phi \in \mathbb{R}$. □

Example (Famille exponentielle gaussienne)

Soit $Y \sim N(\mu, \sigma^2)$. On peut écrire $f(y) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\dfrac{1}{2}\left(\dfrac{y-\mu}{\sigma}\right)^2\right\} =$

$$= \exp\left\{-\frac{1}{2\sigma^2}y^2 + \frac{\mu}{\sigma^2}y - \frac{1}{2}\log(2\pi\sigma^2) - \frac{\mu^2}{2\sigma^2}\right\}.$$

On définit

$$\phi_1 = \frac{\mu}{\sigma^2}, \quad \phi_2 = -\frac{1}{2\sigma^2},$$

$$T_1(y) = y, \quad T_2(y) = y^2, \quad S(y) = 0, \quad \gamma(\phi_1, \phi_2) = -\frac{\phi_1^2}{4\phi_2} + \frac{1}{2}\log\left(-\frac{\pi}{\phi_2}\right),$$

et on remarque que le support de f est toujours $\mathbb{R}$. Donc $N(\mu, \sigma^2)$ est une famille exponentielle à deux paramètres. □

Remarque : Nous concluons une propriété très profonde, à savoir que la loi gaussienne $\mathrm{N}(\mu, \sigma^2)$ possède l'entropie maximale parmi toutes les lois supportées sur $\mathbb{R}$ ayant pour espérance μ et variance σ^2.

Vecteurs aléatoires et lois conjointes

Un vecteur aléatoire $\mathrm{X} = (X_1, \ldots, X_d)^\top$ est une collection finie de variables aléatoires (disposées comme les coordonnées d'un vecteur)

L'idée est que l'on peut vouloir formuler des énoncés probabilistes sur le comportement conjoint de toutes ces variables aléatoires.

- La fonction de répartition conjointe d'un vecteur aléatoire $\mathrm{X} = (X_1, \ldots, X_d)^\top$ est définie par :

$$F_{\mathrm{X}}(x_1, \ldots, x_d) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d).$$

- En correspondance, on définit :
  - la fonction de masse conjointe, si les $\{X_i\}_{i=1}^d$ sont toutes discrètes,

$$f_{\mathrm{X}}(x_1, \ldots, x_d) = \mathbb{P}(X_1 = x_1, \ldots, X_d = x_d).$$

  - la densité conjointe, s'il existe $f_{\mathrm{X}} : \mathbb{R}^d \to [0, +\infty)$ telle que :

$$F_{\mathrm{X}}(x_1, \ldots, x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f_{\mathrm{X}}(u_1, \ldots, u_d)\, du_1 \ldots du_d$$

Dans ce cas, lorsque f_{X} est continue au point $\mathbf{x}$,

$$f_{\mathrm{X}}(x_1, \ldots, x_d) = \frac{\partial^d}{\partial x_1 \ldots \partial x_d} F_{\mathrm{X}}(x_1, \ldots, x_d).$$

Example (Loi conjointe de la somme et du maximum de deux dés)

On lance deux dés équilibrés et indépendants $D_1, D_2 \sim \{1, \dots, 6\}$. On s'intéresse au vecteur aléatoire

$$\mathrm{X} = (S, M), \qquad S = D_1 + D_2, \quad M = \max(D_1, D_2).$$

D'abord, considérons le support de la loi conjointe:

- S prend des valeurs dans $\{2, \dots, 12\}$.
- M prend des valeurs dans $\{1, \dots, 6\}$.
- Mais toutes les paires (s, m) ne sont pas possibles : par exemple $(s = 2, m = 6)$ est impossible.

Et maintenant, les probabilités : chaque issue (d_1, d_2) est équiprobable $(1/36)$.
On compte le nombre d'issues correspondant à chaque couple $(S = s, M = m)$.

$$\mathbb{P}(S = s, M = m) \;=\; \frac{\#\{(d_1, d_2) : d_1 + d_2 = s, \max(d_1, d_2) = m\}}{36}.$$

Par exemple:

- $\mathbb{P}(S = 7, M = 4) = \frac{2}{36}$ car les issues sont $(3, 4), (4, 3)$.
- $\mathbb{P}(S = 12, M = 6) = \frac{1}{36}$ car seule l'issue $(6, 6)$ convient.

Il est commode de disposer les probabilités dans un tableau, où l'entrée (i, j) indique la probabilité que l'issue soit (i, j).

$M \backslash S$	2	3	4	5	6	7	8	9	10	11	12
1	$\frac{1}{36}$										
2		$\frac{2}{36}$	$\frac{1}{36}$								
3			$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$						
4				$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$				
5					$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$		
6						$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Chaque probabilité est obtenue en comptant le nombre d'issues (d_1, d_2) réalisant $(S = s, M = m)$ puis en divisant par 36.

Example (Vecteur aléatoire uniforme sur $S \subset \mathbb{R}^2$)

Soit $S \subset \mathbb{R}^2$ de surface (aire) $a \in (0, \infty)$. Un vecteur aléatoire $X = (X_1, X_2)$ est uniforme sur S s'il admet la densité

$$f(x_1, x_2) = \frac{1}{a} 1_S(x_1, x_2) = \begin{cases} \frac{1}{a}, & (x_1, x_2) \in S, \\ 0, & \text{sinon.} \end{cases}$$

Interprétation géométrique : pour tout borélien $A \subset \mathbb{R}^2$,

$$\mathbb{P}\{X \in A\} = \frac{1}{a} \int_A 1_S(x_1, x_2)\, dx_1\, dx_2 = \frac{\text{aire}(A \cap S)}{\text{aire}(S)}.$$

Prenons le support comme $S = [0,1]^2$, alors $a = 1$. On s'intéresse à $\mathbb{P}\{X_1 \leq X_2\} = \mathbb{P}\{(X_1, X_2) \in A\}$ avec

$$A = \{(x_1, x_2) \subset [0,1]^2 : x_1 \leq x_2\}.$$

Alors,

$$\mathbb{P}\{(X_1, X_2) \in A\} = \int_0^1 \int_{x_1}^1 1\, dx_2\, dx_1 = \int_0^1 (1 - x_1)\, dx_1 = \tfrac{1}{2}.$$

Étant donnée la loi conjointe du vecteur aléatoire $\mathrm{X} = (X_1, \ldots, X_d)^\top$, on peut isoler la loi d'une coordonnée, par exemple X_i.

- Cas discret : la fonction de masse marginale de X_i est donnée par

$$f_{X_i}(x_i) = \mathbb{P}(X_i = x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_d} f_{\mathrm{X}}(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d).$$

- Cas continu : la densité marginale de X_i est donnée par

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathrm{X}}(y_1, \ldots, y_{i-1}, x_i, y_{i+1}, \ldots, y_d)\, dy_1 \ldots dy_{i-1}\, dy_{i+1}\, dy_d.$$

Il s'ensuit que la fonction de repartition marginale de X_i est liée a la fonction de repartition conjointe de $\mathrm{X} = (X_1, \ldots, X_d)^\top$ via

$$F_{X_i}(u) = \lim_{x_j \to \infty, j \neq i} F_{\mathrm{X}}(x_1, ..., x_{i-1}, u, x_{i+1}, ..., x_d)$$

Plus généralement, on peut définir la masse/densité conjointe d'un sous-ensemble des coordonnées :

- par exemple les k premières dans le cas discret :

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \sum_{x_{k+1}} \cdots \sum_{x_d} f_{\mathrm{X}}(x_1,\ldots,x_k,x_{k+1},\ldots,x_d).$$

- ou les k premières dans le cas continu :

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_{\mathrm{X}}(x_1,\ldots,x_k,x_{k+1},\ldots,x_d)\, dx_{k+1}\ldots dx_d.$$

Autrement dit, pour marginaliser on somme/intègre les variables restantes.

Example (Maximum/somme de deux dés – lois marginales)

Rappel : l'entrée (i,j) indique $\mathbb{P}(M=i, S=j)$, où S est la somme et M le maximum.

Les sommes par lignes donnent la marginale de M et les sommes par colonnes celle de S.

$M \backslash S$	2	3	4	5	6	7	8	9	10	11	12	$\sum_s$
1	$\frac{1}{36}$											$\frac{1}{36}$
2		$\frac{2}{36}$	$\frac{1}{36}$									$\frac{3}{36}$
3			$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$							$\frac{5}{36}$
4				$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$					$\frac{7}{36}$
5					$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$			$\frac{9}{36}$
6						$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{11}{36}$
$\sum_m$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

Cela explique le nom "marginales" : ce sont précisément les probabilités que l'on obtient aux marges (lignes/colonnes) du tableau.

Attention : Les marginales ne déterminent pas de façon unique la loi conjointe.

Par exemple, considérons un vecteur aléatoire $X = (X_1, X_2)$ de densité

$$f(x_1, x_2) = \begin{cases} 2 & \text{si } (x_1, x_2) \in [0, \frac{1}{2}]^2 \text{ ou } [\frac{1}{2}, 1]^2, \\ 0 & \text{sinon.} \end{cases}$$

- Marginalement, chaque coordonnée X_1 et X_2 est uniforme sur $[0, 1]$.
- Pourtant, la loi conjointe n'est pas uniforme sur le carré unité : le support est seulement deux sous-carrés.
- Exemple : $\mathbb{P}(X_1 \leq \frac{1}{2}, X_2 > \frac{1}{2}) = 0$, alors que ce serait $\frac{1}{4}$ si la loi était uniforme sur tout $[0, 1]^2$.

- De façon intuitive, connaître le comportement de chaque variable prise isolément ne nous dit pas comment elles interagissent.
- Il faut aussi savoir quelque chose sur leur structure de dépendance.
- Cela se reflète dans la notion de loi conditionnelle.

Il peut être utile de formuler des énoncés probabilistes sur les valeurs possibles d'une variable aléatoire, si l'on connaît déjà la réalisation d'une autre.

Pour cela, on introduit la notion de fonction de densité/de masse conditionnelle.

Si $(X_1, \ldots, X_d)$ est un vecteur aléatoire continu ou discret, on définit la fonction de probabilité conditionnelle de densité/de masse de $(X_1, \ldots, X_k)$ sachant $\{X_{k+1} = x_{k+1}, \ldots, X_d = x_d\}$ par

$$f_{X_1,\ldots,X_k \mid X_{k+1},\ldots,X_d}(x_1, \ldots, x_k \mid x_{k+1}, \ldots, x_d) = \frac{f_{X_1,\ldots,X_d}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_d)}{f_{X_{k+1},\ldots,X_d}(x_{k+1}, \ldots, x_d)}$$

à condition que $f_{X_{k+1},\ldots,X_d}(x_{k+1}, \ldots, x_d) > 0$.

Pare exemple, dans le cas de deux variables, $\mathrm{X} = (X_1, X_2)^\top$,

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}, \quad \text{lorsque } f_{X_2}(x_2) > 0.$$

Ceci nous permet de définir la fonction de répartition conditionnelle,

$$F_{X_1|X_2}(x_1|x_2) = \int_{-\infty}^{x_1} f_{X_1|X_2}(u|x_2)\,du, \quad \text{lorsque } f_{X_2}(x_2) > 0.$$

(et le cas générale par analogie)

Example

$$\text{On revient sur} \quad f(x_1, x_2) = \begin{cases} 2, & (x_1, x_2) \in [0, \frac{1}{2}]^2 \text{ ou } [\frac{1}{2}, 1]^2, \\ 0, & \text{sinon.} \end{cases}$$

Vérification : chaque carré a aire $1/4$, donc masse $2 \times 1/4 = 1/2$. La somme donne 1 : c'est bien une densité. Les marginales sont

$$f_{X_1}(x_1) = \int_0^1 f(x_1, x_2)\, dx_2 = \begin{cases} 1, & x_1 \in [0, 1], \\ 0, & \text{sinon,} \end{cases}$$

et de même

$$f_{X_2}(x_2) = 1, \qquad x_2 \in [0, 1].$$

Les conditionnelles se lisent directement :

$$f_{X_2|X_1}(x_2|x_1) = \begin{cases} 2, & (x_1, x_2) \in [0, \frac{1}{2}]^2 \text{ ou } [\frac{1}{2}, 1]^2, \\ 0, & \text{sinon,} \end{cases}$$

c'est-à-dire : sachant X_1, X_2 est uniforme sur le même intervalle demi-unitaire (l'autre s'obtient de façon symétrique).

Example

On considère la densité jointe

$$f_{X,Y}(x,y) = c\, e^{-x-y}\, 1(y > x)\, 1(x > 0).$$

La densité marginale de X est $f_X(x) = c \int_{y=x}^{\infty} e^{-x-y}\, dy = ce^{-2x}$, $x > 0$. Cette fonction s'intègre à 1 seulement si $c = 2$. La densité marginale de Y est

$$f_Y(y) = 2 \int_{x=0}^{y} e^{-x-y}\, dx = 2e^{-y}(1 - e^{-y}), \qquad y > 0,$$

et $\int_0^{\infty} f_Y(y)\, dy = 1$, donc c'est bien une densité. Alors

$$f(y|x) = \frac{2e^{-x-y}}{2e^{-2x}} = e^{x-y}, \qquad y > x,$$

et

$$f(x|y) = \frac{2e^{-x-y}}{2e^{-y}(1 - e^{-y})} = \frac{e^{-x}}{1 - e^{-y}}, \qquad 0 < x < y.$$

On vérifie facilement que ces deux densités s'intègrent à 1.

Les variables aléatoires $X_1, \ldots, X_d$ sont dites indépendantes si et seulement si, pour tout $x_1, \ldots, x_d \in \mathbb{R}$,

$$F_{X_1,\ldots,X_d}(x_1, \ldots, x_d) \;=\; F_{X_1}(x_1) \times \ldots \times F_{X_d}(x_d).$$

De manière équivalente, $X_1, \ldots, X_d$ sont indépendantes si et seulement si, pour tout $x_1, \ldots, x_d \in \mathbb{R}$,

$$f_{X_1,\ldots,X_d}(x_1, \ldots, x_d) \;=\; f_{X_1}(x_1) \times \ldots \times f_{X_d}(x_d).$$

Pour deux variables aléatoires X et Y, on note leur indépendance $X \perp\!\!\!\perp Y$.

À noter que lorsque des variables aléatoires sont indépendantes, les lois conditionnelles se réduisent aux lois marginales correspondantes (exercice).

Sous l'indépendance, connaître la valeur de l'une des variables aléatoires ne nous donne aucune information sur la loi des autres.

Example (Lancers de pièces indépendants)

Soit $X_1, \ldots, X_n$ une suite de variables aléatoires indépendantes $\mathrm{Bern}(p)$,

$$\mathbb{P}(X_i = 1) = p, \qquad \mathbb{P}(X_i = 0) = 1 - p.$$

La probabilité d'une configuration particulière $\mathbf{x} = (x_1, \ldots, x_n) \in \{0,1\}^n$ est

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} = f_{\mathbf{X}}(\mathbf{x}).$$

Si l'on s'intéresse uniquement au nombre total de succès $S_n = \sum_{i=1}^n X_i$, alors

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Différence essentielle :

- dans la loi jointe ci-dessus, on considère une configuration particulière (un *permutation* de k succès et $n-k$ échecs) ;
- dans la loi binomiale, on regroupe toutes les configurations menant à k succès (une *combinaison*) – les événements liés à configurations distincts sont disjoints.

Le cas discret ne change pas en ce qui concerne les transformations de variables : on peut définir une variable discrète scalaire en énumérant tous les résultats — ils sont dénombrables — et appliquer les résultats connus pour le cas scalaire.

Pour le cas continu : soit $g : \mathbb{R}^n \to \mathbb{R}^n$ une bijection différentiable,

$$g(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_n(\mathbf{x})), \qquad \mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n.$$

Soit $X = (X_1, \ldots, X_n)^\top$ de densité conjointe $f_{\mathrm{X}}(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$, et posons $\mathrm{Y} = (Y_1, \ldots, Y_n)^\top = g(\mathrm{X})$. Alors, Y prend ses valeurs dans $\mathcal{Y}^n = g(\mathcal{X}^n)$, et

$$f_{\mathrm{Y}}(\mathbf{y}) = f_{\mathrm{X}}(g^{-1}(\mathbf{y}))\left|\det\left[J_{g^{-1}}(\mathbf{y})\right]\right|, \qquad \text{pour } \mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathcal{Y}^n,$$

et vaut zéro sinon, lorsque $J_{g^{-1}}(\mathbf{y})$ est bien défini. Ici, $J_{g^{-1}}(\mathbf{y})$ est le jacobien de g^{-1}, c'est-à-dire la fonction matricielle $n \times n$,

$$J_{g^{-1}}(\mathbf{y}) = \begin{bmatrix} \frac{\partial}{\partial y_1} g_1^{-1}(\mathbf{y}) & \cdots & \frac{\partial}{\partial y_n} g_1^{-1}(\mathbf{y}) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial y_1} g_n^{-1}(\mathbf{y}) & \cdots & \frac{\partial}{\partial y_n} g_n^{-1}(\mathbf{y}) \end{bmatrix}.$$

Preuve.

Soit $A \subset \mathcal{Y}^n$. Alors

$$\mathbb{P}\{\mathrm{Y} \in A\} = \mathbb{P}\{g(\mathrm{X}) \in A\} = \mathbb{P}\{\mathrm{X} \in g^{-1}(A)\}.$$

Par définition de la densité de X,

$$\mathbb{P}\{\mathrm{X} \in g^{-1}(A)\} = \int_{g^{-1}(A)} f_{\mathrm{X}}(\mathrm{x})\, d\mathrm{x}.$$

Comme g est bijective, on peut effectuer le changement de variables $\mathrm{x} = g^{-1}(\mathrm{y})$, ce qui donne

$$\int_{g^{-1}(A)} f_{\mathrm{X}}(\mathrm{x})\, d\mathrm{x} = \int_A f_{\mathrm{X}}(g^{-1}(\mathrm{y}))\, \left|\det J_{g^{-1}}(\mathrm{y})\right|\, d\mathrm{y}.$$

Ainsi, la densité de Y est

$$f_{\mathrm{Y}}(\mathrm{y}) = f_{\mathrm{X}}(g^{-1}(\mathrm{y}))\, \left|\det J_{g^{-1}}(\mathrm{y})\right|, \qquad \mathrm{y} \in \mathcal{Y}^n.$$

□

Example (Loi normale bivariée)

Soient Z_1 et Z_2 deux variables aléatoires indépendantes $\mathrm{N}(0,1)$. Par indépendance, leur densité conjointe est

$$f_{\mathrm{Z}}(z_1, z_2) = \frac{1}{2\pi} \exp\left\{-\tfrac{1}{2}(z_1^2 + z_2^2)\right\}, \quad (z_1, z_2)^\top \in \mathbb{R}^2,$$

On définit maintenant une transformation linéaire inversible $\mathrm{z} \mapsto \mathrm{L}z + \mu$,

$$\mu = (\mu_1, \mu_2)^\top \in \mathbb{R}^2, \quad \mathrm{L} = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2\sqrt{1-\rho^2} \end{pmatrix}, \quad \mathrm{L}^{-1} = \begin{pmatrix} \frac{1}{\sigma_1} & 0 \\ -\frac{\rho}{\sigma_1\sqrt{1-\rho^2}} & \frac{1}{\sigma_2\sqrt{1-\rho^2}} \end{pmatrix}.$$

L'application inverse est $\mathrm{x} \mapsto \mathrm{L}^{-1}(\mathrm{x} - \mu)$, on obtient que $\mathrm{X} = \mathrm{LZ}$ a pour densité

$$f_{\mathrm{X}}(\mathrm{x}) = \frac{1}{2\pi|\det(\Sigma)|^{1/2}} \exp\left\{-\tfrac{1}{2}(\mathrm{x}-\mu)^\top \Sigma^{-1}(\mathrm{x}-\mu)\right\}$$

$$= \frac{1}{2\pi\,\sigma_1\,\sigma_2\,\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho\,(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\},$$

où

$$\Sigma = \mathrm{L}\mathrm{L}^\top = \begin{pmatrix} \sigma_1^2 & \rho\,\sigma_1\sigma_2 \\ \rho\,\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

Example (Convolution de densités)

Soient X et Y deux variables aléatoires continues et indépendantes de densités f_X et f_Y. La densité de $X + Y$ est la *convolution* de f_X avec f_Y :

$$f_{X+Y}(u) = \int_{-\infty}^{+\infty} f_X(u-v) f_Y(v)\, dv.$$

Définissons

$$g : \mathbb{R}^2 \to \mathbb{R}^2, \qquad (x,y) \overset{g}{\mapsto} (x+y, y) \qquad (u,v) \overset{g^{-1}}{\mapsto} (u-v, v).$$

Le jacobien de l'inverse est

$$\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$$

et son déterminant vaut 1. Il en découle que

$$f_{X+Y,Y}(u,v) = f_{X,Y}(u-v, v) = f_X(u-v) f_Y(v),$$

et on intègre par rapport à v pour trouver la marginale f_{X+Y} :

$$f_{X+Y}(u) = \int_{-\infty}^{+\infty} f_X(u-v) f_Y(v)\, dv.$$

Moments produits et leurs génératrices

Soit $\mathrm{X} = (X_1, \ldots, X_d)^\top$ un vecteur aléatoire dans $\mathbb{R}^d$ de densité conjointe $f_{\mathrm{X}}(x_1, \ldots, x_d)$. Pour toute application $g : \mathbb{R}^d \to \mathbb{R}$, on définit

$$\mathbb{E}\{g(X_1, \ldots, X_d)\} = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} g(x_1, \ldots, x_d) f_{\mathrm{X}}(x_1, \ldots, x_d)\, dx_1 \ldots dx_d.$$

De même, dans le cas discret,

$$\mathbb{E}\{g(X_1, \ldots, X_d)\} = \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_d \in \mathcal{X}_d} g(x_1, \ldots, x_d) f_{\mathrm{X}}(x_1, \ldots, x_d).$$

Le vecteur moyen d'un vecteur aléatoire $\mathrm{X} = (X_1, \ldots, X_d)$ est défini par

$$\mathbb{E}[\mathrm{X}] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_d] \end{pmatrix},$$

c'est-à-dire le vecteur des espérances.

Quelle que soit la loi conjointe de (X_1, X_2), l'espérance de leur somme est la somme des espérances (donc déterminée par les lois marginales) — Attention : l'espérance du produit, par contre, n'est pas déterminée par les lois marginales, comme on le verra plus tard.

Soit (X_1, X_2) un vecteur aléatoire continu de densité conjointe $f_{X_1,X_2}(x_1, x_2)$. Pour $a \in \mathbb{R}$, considérons $aX_1 + X_2$.

$$\begin{aligned}
\mathbb{E}[aX_1 + X_2] &= \iint_{\mathbb{R}^2} (ax_1 + x_2)\, f_{X_1,X_2}(x_1, x_2)\, dx_1\, dx_2 \\
&= a \iint_{\mathbb{R}^2} x_1 f_{X_1,X_2}(x_1, x_2)\, dx_1\, dx_2 + \iint_{\mathbb{R}^2} x_2 f_{X_1,X_2}(x_1, x_2)\, dx_1\, dx_2 \\
&= a\,\mathbb{E}[X_1] + \mathbb{E}[X_2].
\end{aligned}$$

$$\Longrightarrow \quad \mathbb{E}[aX_1 + X_2] = a\,\mathbb{E}[X_1] + \mathbb{E}[X_2].$$

(pour le cas discret, on remplace les intégrales par des sommes)

Corollaire : pour tout vecteur aléatoire $\mathrm{X} = (X_1, ..., X_d)^\top$ on a $\mathbb{E}[\mathrm{AX}] = \mathrm{A}\mathbb{E}[\mathrm{X}]$ pour toute matrice deterministe $\mathrm{A}_{n\times d}$.

Rappelons que la variance d'une variable aléatoire X exprime la dispersion des réalisations de X autour de son espérance.

$$\mathrm{var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] \qquad (\text{si } \mathbb{E}[X^2] < \infty).$$

Dans le cas vectoriel, la covariance d'une variable aléatoire X_1 avec une autre variable X_2 exprime le degré de dépendance linéaire entre les deux.

$$\mathrm{cov}(X_1, X_2) = \mathbb{E}\left[(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))\right] \qquad (\text{si } \mathbb{E}[X_i^2] < \infty).$$

La corrélation entre X_1 et X_2 est définie par

$$\mathrm{Corr}(X_1, X_2) = \frac{\mathrm{cov}(X_1, X_2)}{\sqrt{\mathrm{var}(X_1)\mathrm{var}(X_2)}}.$$

Elle transmet une information de dépendance équivalente à la covariance. Avantages : (1) elle est invariante par changement d'échelle, (2) elle s'interprète en valeur absolue (bornée par $[-1, 1]$), grâce à l'inégalité de corrélation :

$$|\mathrm{Corr}(X_1, X_2)| \leq 1.$$

(elle-même conséquence de l'inégalité de Cauchy–Schwarz)

Quelques formules utiles reliant espérances, variances et covariances :

- $\mathrm{cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2]$
- $\mathrm{cov}(X, X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathrm{var}(X)$
- $\mathrm{cov}(aX_1 + bX_2, cY) = ac\,\mathrm{cov}(X_1, Y) + bc\,\mathrm{cov}(X_2, Y)$ (bilinearité)
- (alors on retrouve) $\mathrm{var}(aX + b) = a^2\mathrm{var}(X)$
- $\mathrm{var}(\sum_i X_i) = \sum_i \mathrm{var}(X_i) + \sum_{i \neq j} \mathrm{cov}(X_i, X_j)$
- si $\mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] < \infty$, alors les propriétés suivantes sont équivalentes :

  (i) $\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2]$

  (ii) $\mathrm{cov}(X_1, X_2) = 0$

  (iii) $\mathrm{var}(X_1 \pm X_2) = \mathrm{var}(X_1) + \mathrm{var}(X_2)$

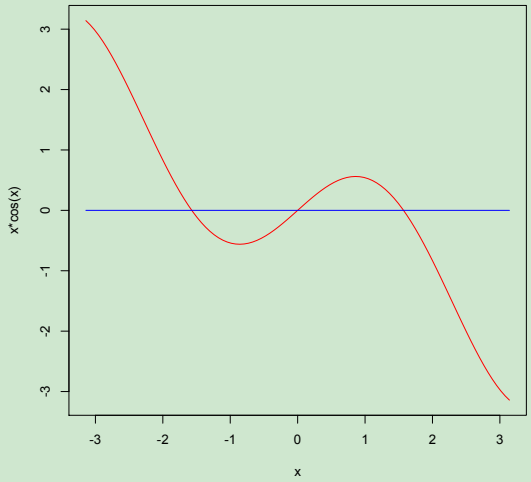
  L'indépendance implique ces trois dernières propriétés, mais aucune d'entre elles n'implique l'indépendance.

Example ($\mathrm{Corr}(X, Y) = 0 \not\Rightarrow$ Indépendance)

Soit $X \sim \mathrm{Unif}[-\pi, \pi]$ et définissons

$$Y = \cos(X).$$

- Clairement, X et Y ne sont pas indépendantes.
- Au contraire, elles sont parfaitement dépendantes.
- Leur covariance est néanmoins nulle !



La fonction $x \cos(x)$

Concrètement, on calcule

$$\mathbb{P}[Y > 0] = 1/2 \qquad \text{mais} \qquad \mathbb{P}[Y > 0 \,|\, X \in (-\pi, -2)] = 1.$$

Malgré cela, on a

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \int_{-\pi}^{+\pi} x \cos(x) \frac{1}{2\pi}\, dx - 0 = 0.$$

Pourquoi : la covariance est aveugle à de dépendances "purement non linéaires" (pente zero)

Example ($\mathrm{Corr}(X, Y) = 0 \not\Rightarrow$ Indépendance)

Soient X et Y de densité conjointe

$$f_{XY}(x, y) = \begin{cases} 1/\pi & \text{si } x^2 + y^2 \leq 1, \\ 0 & \text{sinon.} \end{cases}$$

On note que $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ par symétrie. Ainsi, $\mathrm{Cov}(X, Y) = \mathbb{E}[XY]$. Or

$$\mathbb{E}[XY] = \iint_{x^2+y^2\leq 1} xy \frac{1}{\pi}\, dxdy = \iint_{x^2+y^2\leq 1,\, y\geq 0} xy \frac{1}{\pi}\, dxdy + \iint_{x^2+y^2\leq 1,\, y<0} xy \frac{1}{\pi}\, dxdy.$$

Les deux termes sont égaux par symétrie. De plus,

$$\iint_{x^2+y^2\leq 1,\, y\geq 0} xy \frac{1}{\pi}\, dxdy = \frac{1}{\pi} \int_{-1}^{1} x \int_{0}^{\sqrt{1-x^2}} y\, dy\, dx = \frac{1}{\pi} \int_{-1}^{1} x \frac{(1-x^2)}{2}\, dx = 0,$$

et donc la corrélation est nulle. Mais X et Y sont clairement dépendantes, puisque connaître X restreint les valeurs possibles de Y.

Pourquoi : l'independence nécessite un support rectangulaire, mais on peut avoir covariance zero même avec support non-rectangulaire

La matrice de covariance d'un vecteur aléatoire $\mathrm{X} = (X_1, \ldots, X_d)^\top$, notée $\Sigma = \{\Sigma_{ij}\}$, est une matrice $d \times d$ symétrique avec

$$\Sigma_{ij} = \mathrm{cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])], \quad 1 \leq i \leq j \leq d.$$

Autrement dit, la matrice de covariance contient les variances des coordonnées de X (sur la diagonale) et les covariances entre toutes paires de coordonnées de X (hors diagonale).

Si l'on note

$$\mu = \mathbb{E}[\mathrm{X}] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])^\top$$

le vecteur moyen de X, alors

$$\Sigma = \mathbb{E}[(\mathrm{X} - \mu)(\mathrm{X} - \mu)^\top] = \mathbb{E}[\mathrm{X}\mathrm{X}^\top] - \mu\mu^\top.$$

Comme dans le cas vectoriel, l'espérance d'une matrice à entrées aléatoires est la matrice des espérances de ses entrées.

Soit X un vecteur aléatoire $d \times 1$ de vecteur moyen μ et de matrice de covariance Σ. Grâce à la linéarité de l'espérance et à la bilinéarité de la covariance, on a

$$\begin{aligned} \mathrm{cov}\{\beta^\top \mathrm{X}, \gamma^\top \mathrm{X}\} &= \mathbb{E}[\beta^\top \mathrm{X}\mathrm{X}^\top \gamma] - \mathbb{E}[\beta^\top \mathrm{X}]\mathbb{E}[X^\top \gamma] = \beta^\top \big(\mathbb{E}[\mathrm{X}\mathrm{X}^\top] - \mu\mu^\top\big)\gamma \\ &= \beta^\top \Sigma \gamma \end{aligned}$$

pour tous vecteurs $\beta, \gamma \in \mathbb{R}^d$ déterministes. Il s'ensuit que :

- Si $\beta \in \mathbb{R}^d$ est un vecteur déterministe, la variance de $\beta^\top \mathrm{X}$ est $\beta^\top \Sigma \beta$.
- Pour tout $\beta \in \mathbb{R}^d$, on a $\beta^\top \Sigma \beta \geq 0$.
- Si A est une matrice déterministe $p \times d$, alors le vecteur moyen et la matrice de covariance de AX sont $\mathrm{A}\mu$ et $\mathrm{A}\Sigma\mathrm{A}^\top$, respectivement.

De même que les moments d'ordre supérieur donnent des caractéristiques plus fines de la loi (marginale) d'une variable aléatoire X, les moments produits d'ordre supérieur impliquant les coordonnées d'un vecteur aléatoire reflètent des dépendances plus fines entre ces coordonnées.

- La covariance concerne des paires — elle reflète des dépendances *bilatérales*.
- Si l'on veut formuler des énoncés de dépendance "d'ordre supérieur", c.-à-d. concernant un "tuplet" $(X_1, \ldots, X_d)$, il faut examiner des moments du type :

$$\mathbb{E}\left[\prod_{j=1}^{d} X_j^{k_j}\right] = \mathbb{E}\left[X_1^{k_1} \times \ldots \times X_d^{k_d}\right].$$

[*le rôle de ces moments produits s'apprécie en considérant comment une fonction (lisse) générale* $g(X_1, \ldots, X_d)$ *peut être approchée par polynômes en* $(X_1, \ldots, X_d)$ *(série de Taylor)*].

- On peut alors se demander : si ces moments produits expriment des dépendances de plus en plus fines, est-ce que les connaître tous détermine la loi conjointe de façon unique ?
- La réponse est similaire au cas marginal, et fait intervenir la notion de fonction génératrice des moments d'un vecteur aléatoire.

La fonction génératrice des moments (FGM) d'un vecteur aléatoire $\mathrm{X} = (X_1, \ldots, X_d)^\top$ est définie par

$$M_{\mathrm{X}}(\mathrm{t}) = \mathbb{E}\left[e^{\mathrm{t}^\top \mathrm{X}}\right] = \mathbb{E}\left[e^{t_1 X_1 + \ldots + t_d X_d}\right], \qquad \mathrm{t} = (t_1, \ldots, t_d)^\top \in \mathbb{R}^d.$$

Elle n'est pas forcément finie pour $\mathrm{t} \neq 0$. Mais lorsqu'elle est finie dans un voisinage de l'origine, plusieurs propriétés importantes en découlent :

1. M_{X} est indéfiniment différentiable dans ce voisinage.
2. Les moments existent : pour tout multi-indice $\mathrm{k} = (k_1, \ldots, k_d)$,

$$\mathbb{E}[X_1^{k_1} \cdots X_d^{k_d}] = \frac{\partial^{|\mathrm{k}|} M_{\mathrm{X}}}{\partial t_1^{k_1} \cdots \partial t_d^{k_d}}(0).$$

3. La loi conjointe de X est entièrement déterminée par M_{X} (si elle est finie autour de l'origine).
4. Si X et Y sont indépendants, alors

$$M_{\mathrm{X}+\mathrm{Y}}(\mathrm{t}) = M_{\mathrm{X}}(\mathrm{t})\, M_{\mathrm{Y}}(\mathrm{t}).$$

Example (MGF d'une loi binomiale)

Soit $X_1, \ldots, X_n$ une suite de variables aléatoires indépendantes, chacune suivant une loi de Bernoulli(p), et posons

$$S_n = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p).$$

L'MGF d'une Bernoulli(p) est

$$M_X(t) = \mathbb{E}[e^{tX}] = (1-p)\, e^{t\cdot 0} + p\, e^{t\cdot 1} = 1 - p + pe^t.$$

Par indépendance,

$$M_{S_n}(t) = \mathbb{E}[e^{t\sum_{i=1}^{n} X_i}] = \prod_{i=1}^{n} \mathbb{E}[e^{tX_i}] = \Big(M_X(t)\Big)^n.$$

Donc

$$M_{S_n}(t) = (1 - p + pe^t)^n.$$

Example (Moments centraux supérieurs d'une Gaussienne bivariée)

Soit $\mathrm{X} = (X_1, X_2)^\top \sim \mathrm{N}(\mu, \Sigma)$ avec $\mu = (\mu_1, \mu_2)^\top$ et $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$.

Posons la version centrée $\mathrm{U} = \mathrm{X} - \mu$. Écrivons $\Sigma = LL^\top$ comme avant et soit $\mathrm{Z} \sim \mathrm{N}(0, \mathrm{I})$. Alors $\mathrm{U} = L\mathrm{Z}$ et, pour tout $\mathrm{t} \in \mathbb{R}^2$,

$$M_{\mathrm{U}}(\mathrm{t}) = \mathbb{E}\Big[e^{\mathrm{t}^\top \mathrm{U}}\Big] = \mathbb{E}\Big[e^{(L^\top \mathrm{t})^\top \mathrm{Z}}\Big] = \exp\Big(\tfrac{1}{2}\|L^\top \mathrm{t}\|^2\Big) = \exp\Big(\tfrac{1}{2}\,\mathrm{t}^\top \Sigma\,\mathrm{t}\Big).$$

La fonction $M_{\mathrm{U}}(\mathrm{t}) = \exp\big(\frac{1}{2}\,\mathrm{t}^\top \Sigma\,\mathrm{t}\big)$ est *paire* en t, donc toutes ses dérivées partielles d'ordre total impair en $\mathrm{t} = 0$ sont nulles. Par définition des moments via dérivées en 0,

$$\mathbb{E}\big[(X_1 - \mu_1)^k\,(X_2 - \mu_2)^m\big] = 0 \quad \text{lorsque } k + m \text{ est impair.}$$

Remarque (ordre pair). Pour $k + m$ pair, les moments se déduisent par dérivation de M_{U} (formule d'Isserlis/Wick), par ex. :

$$\mathbb{E}\big[(X_1 - \mu_1)^2 (X_2 - \mu_2)^2\big] = \sigma_1^2\sigma_2^2 + 2\,\mathrm{cov}(X_1, X_2)^2.$$

On constate que tous les moments d'ordre pair dépendent uniquement de la covariance — la structure de dépendance est fondamentalement linéaire.

On peut calculer l'espérance conditionnelle d'une variable aléatoire X sachant qu'une autre variable aléatoire Y a pris la valeur y par

$$\mathbb{E}[X \,|\, Y = y] = \begin{cases} \sum_{x \in \mathcal{X}} x \, \mathbb{P}[X = x \,|\, Y = y], & \text{si } X, Y \text{ sont discrètes,} \\ \int_{-\infty}^{+\infty} x \, f_{X|Y}(x|y) \, dx, & \text{si } X, Y \text{ sont continues.} \end{cases}$$

- C'est simplement l'espérance de la loi conditionnelle.
- Le calcul de $\mathbb{E}[X \,|\, Y = y] = q(y)$ donne une fonction de y.
- On peut remplacer y par Y et considérer $Z = q(Y)$ comme une variable aléatoire en soi.
- Cette variable est notée $\mathbb{E}[X \,|\, Y]$: c'est la définition formelle de l'espérance conditionnelle.
- Propriété/interprétation importante :

$$\mathbb{E}[X \,|\, Y] = \arg\min_{g} \mathbb{E}\big\|X - g(Y)\big\|^2$$

Parmi toutes les fonctions mesurables de Y, $\mathbb{E}[X \,|\, Y]$ est celle qui approxime le mieux X au sens quadratique moyen.

Example (Espérances conditionnelles)

Rappelons l'exemple

$$f_{X,Y}(x,y) = c\, e^{-x-y}\, 1(y>x)\, 1(x>0).$$

où la densité conditionnelle de Y sachant X est

$$f_{Y|X}(y \mid x) = e^{x-y}, \qquad y > x.$$

Alors,

$$\mathbb{E}[\,Y \mid X = x] = \int_x^\infty y e^{x-y}\, dy = \int_0^\infty (x+u) e^{-u}\, du = x+1, \qquad x>0.$$

De même,

$$\mathbb{E}[\,Y^2 \mid X = x] = \int_x^\infty y^2\, e^{x-y}\, dy = \int_0^\infty (x+u)^2\, e^{-u}\, du = x^2+2x+2, \qquad x>0.$$

Ainsi,

$$\mathbb{E}[\,Y \mid X] = X+1$$

Vecteurs Gaussians (loi multivariée normale)

La loi normale multivariée est sans doute la loi la plus importante en probabilité et en statistique. Elle jouera un rôle central dans la suite du cours. Pourquoi ?

- Maximum d'entropie : parmi toutes les lois sur $\mathbb{R}^d$ ayant même moyenne et covariance, la normale maximise l'entropie.
- Théorème centrale limite : de nombreuses sommes de variables aléatoires tendent vers une loi normale.
- Stabilité : les marges, combinaisons linéaires, et conditionelles de normales restent normales.
- Simplicité analytique : densité explicite, moments faciles à calculer, entièrement décrite par moyenne et covariance.
- Applications : omniprésente en statistique, physique, économie, ingénierie.

Mais cela ne veut pas dire qu'il n'existe pas d'autres lois multivariées importantes. En particulier, certaines lois discrètes jouent un rôle fondamental dans la modélisation.

- Nous verrons rapidement deux exemples instructifs de lois discrètes :
  - la loi multinomiale, qui généralise la binomiale,
  - le modèle d'Ising, un modèle de dépendance en physique statistique.
- Puis nous consacrerons une étude plus détaillée à la loi normale multivariée.

Un vecteur aléatoire $\mathrm{X} \in \mathbb{R}^k$ suit une loi multinomiale de paramètres $n \in \mathbb{N}$ et $p = (p_1, \ldots, p_k) \in (0,1)^k$, avec $\sum_{i=1}^k p_i = 1$, notée $\mathrm{X} \sim \mathrm{Multi}(n; p_1, \ldots, p_k)$, si

1. le support consiste des $\mathrm{x} \in \{0, 1, \ldots, n\}^k$ tels que $x_1 + \ldots + x_k = n$.
2. $$f_{\mathrm{X}}(x_1, \ldots, x_k) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k} \, 1\left\{\sum_{i=1}^k x_i = n\right\}.$$

Les espérances, variances-covariances et la fonction génératrice des moments sont

$$\mathbb{E}[X_i] = np_i, \qquad \mathrm{var}[X_i] = np_i(1-p_i), \qquad \mathrm{Cov}(X_i, X_j) = -np_ip_j,$$

$$M_{\mathrm{X}}(u_1, \ldots, u_k) = \left(\sum_{i=1}^k p_i \, e^{u_i}\right)^n.$$

Lemme (Poisson et multinomiale)

Si $X_i \sim \mathrm{Poiss}(\lambda_i)$, $i = 1, \ldots, k$, sont indépendantes, alors la loi conditionnelle de $\mathrm{X} = (X_1, \ldots, X_k)^\top$ sachant $\sum_{i=1}^k X_i = n$ est $\mathrm{Multi}(n; p_1, \ldots, p_k)$, avec

$$p_i = \frac{\lambda_i}{\lambda_1 + \ldots + \lambda_k}.$$

Points simples mais instructifs sur la loi multinomiale :

- Généralisation de la loi binomiale : La multinomiale décrit n épreuves indépendantes, chacune donnant un résultat parmi k catégories. $\Rightarrow$ la binomiale correspond simplement au cas $k = 2$.

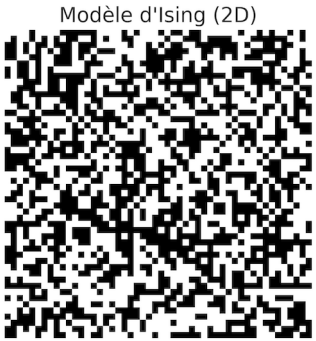
- Lien avec le comptage de configurations : La probabilité d'observer $x_1, \ldots, x_k$ est proportionnelle au nombre de façons de répartir n essais dans k catégories. $\Rightarrow$ d'où le facteur combinatoire $\frac{n!}{x_1! \cdots x_k!}$.

- Correlation négative : $\mathrm{Cov}(X_i, X_j) = -np_ip_j$. $\Rightarrow$ les coordonnées ne sont pas indépendantes, car leur somme vaut n. Des grandes valeurs d'une coordonée sont associés typiquement a des petites valeurs d'une autre.

- Vers une partition de l'unité : Si l'on renormalise par n,

$$\left(\frac{X_1}{n}, \ldots, \frac{X_k}{n}\right),$$

on obtient les proportions dans chaque catégorie. Lorsque $n \to \infty$, les configurations possibles pour ces proportions peuvent être vues comme une partition aléatoire de l'unité via une grille de plus en plus fine.



Un modèle pour le ferromagnétisme, défini sur une grille $\mathcal{G}$ de variables aléatoires $X_j \in \{-1, +1\}$. Donc $\mathcal{X} = \{-1, +1\}^{|\mathcal{G}|}$ a $2^{|\mathcal{G}|}$ éléments, les "configurations".

La fonction de masse conjointe est

$$f_{\mathrm{X}}(\mathbf{x}) = Z(\beta, \mu)^{-1} \exp\left\{\beta\left(\sum_{i\sim j} X_i X_j + \mu \sum_j h_j X_j\right)\right\},$$

où $i \sim j$ ssi X_i et X_j sont adjacents, les h_j correspondent à un champ magnétique externe, β est la température inverse, et $Z(\beta, \mu)$ est la partition function qui normalise la fonction de masse.

- Symétrie sans champ externe : Si $\mu = 0$ (pas de champ externe), le modèle est symétrique par inversion des spins : si $(X_j)_j$ est une configuration, alors $(-X_j)_j$ a la même probabilité. $\Rightarrow$ invariance probabiliste.
- Comportement haute vs. basse température :
  - $\beta \approx 0$ (température très haute) : le facteur exponentiel est proche de 1, donc les $2^{|\mathcal{G}|}$ configurations sont presque équiprobables (comme des lancers indépendants de pièces ± 1).
  - $\beta \to \infty$ (température très basse) : le système favorise fortement les deux états parfaitement ordonnés (tous $+1$ ou tous -1).

  $\Rightarrow$ intuition claire de "désordre vs. ordre".
- Corrélation entre voisins : Pour $\beta = 0$, X_i et X_j sont indépendants. Mais si $\beta > 0$, les voisins sont positivement corrélés : $\mathbb{E}[X_i X_j] > 0$. $\Rightarrow$ "les voisins veulent s'aligner".
- Fonction de partition comme normalisation : Même sans la calculer, $Z(\beta, \mu)$ est ce qui rend la loi jointe une probabilité. Dans de petites grilles, on peut la calculer explicitement en sommant sur un nombre fini de configurations.

Definition (Loi gaussienne multivariée)

Un vecteur aléatoire $Y \in \mathbb{R}^d$ est dit gaussien si et seulement si $\beta^\top Y$ est une variable aléatoire gaussienne pour tout vecteur déterministe $\beta \in \mathbb{R}^d$.

Observation : D'après la définition, Y possède forcément un vecteur moyen vecteur moyen μ et une matrice de covariance Σ bien définis, qui se déduisent alors (de manière unique) en écrivant

$$\mu_i = \mathbb{E}[\mathrm{e}_i^\top Y], \qquad \Sigma_{ij} = \mathrm{cov}\{\mathrm{e}_i^\top Y, \mathrm{e}_j^\top Y\},$$

où e_j est le j-ième vecteur de la base canonique

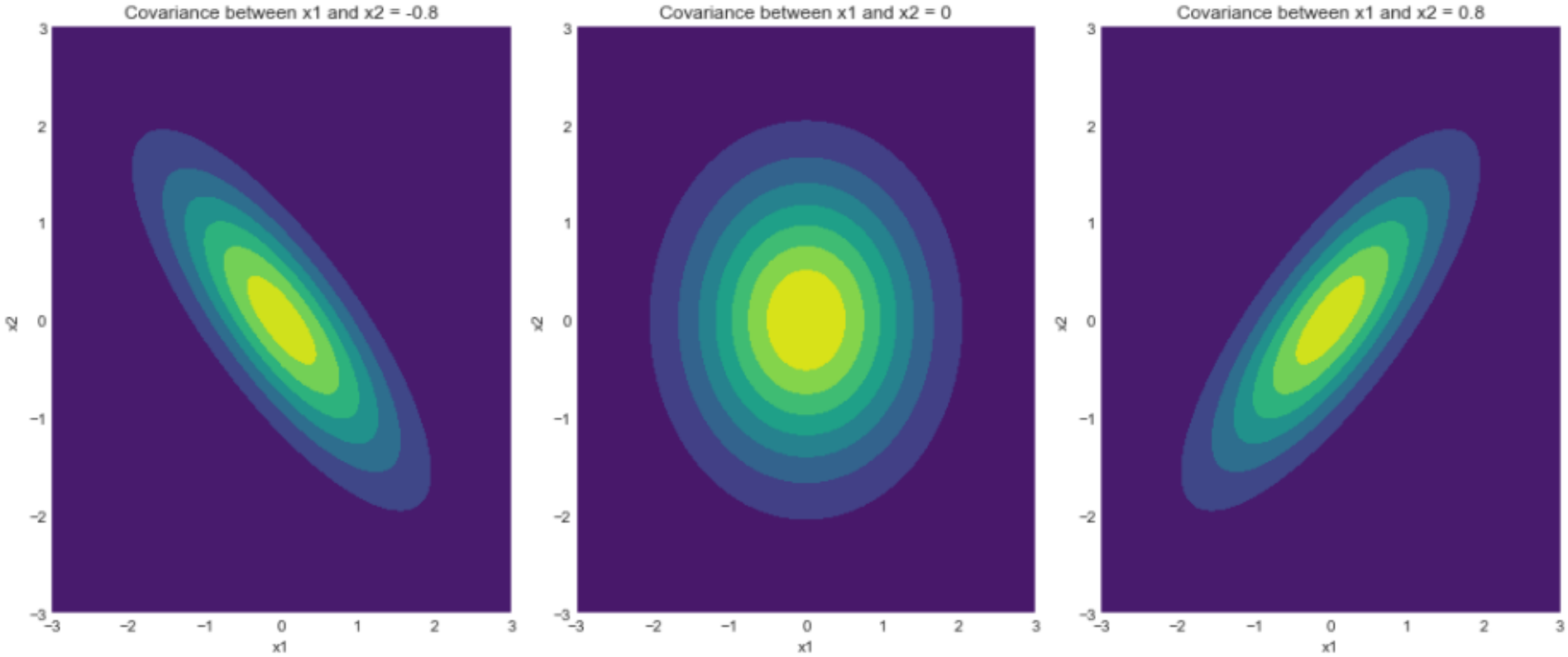
$$\mathrm{e}_j = (0, 0, \ldots, \underbrace{1}_{j\text{-ième position}}, \ldots, 0, 0)^\top.$$

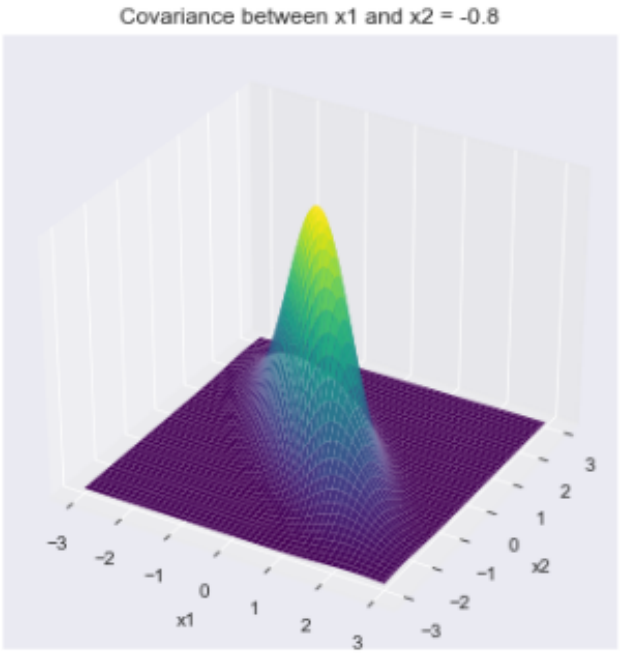
En effet, comme $\mathbb{E}\{(\beta^\top Y)^2\} < \infty$ pour tout β, $\mathbb{E}(e_i^\top Y)^2 < \infty$ pour tout i.

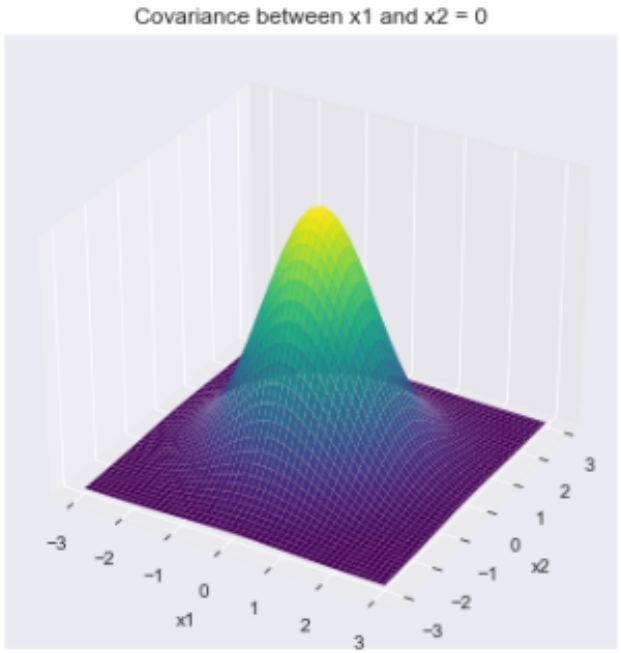
On peut, donc écrire $\mathrm{N}(\mu, \Sigma)$ pour la loi multivariée normale correspondante.

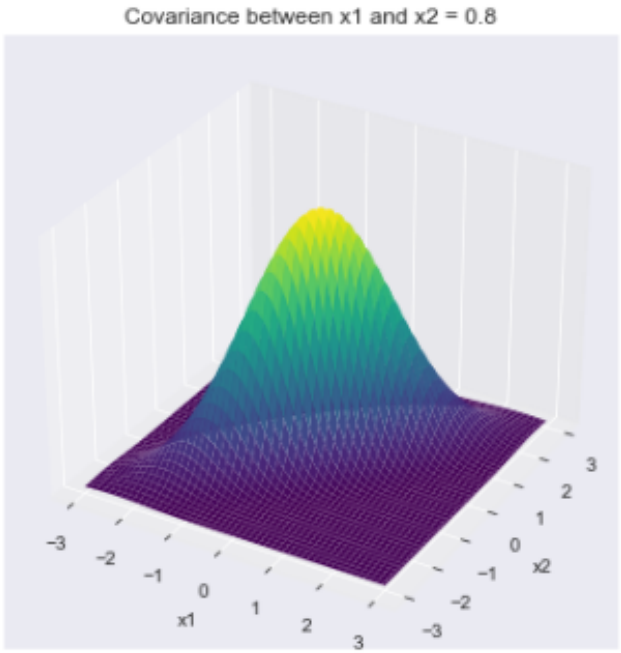
Remarque : par convention, nous considerons tout vecteur deterministe $v \in \mathbb{R}^d$ come ayant une loi normale de moyenne v et de matrice de covariance 0.

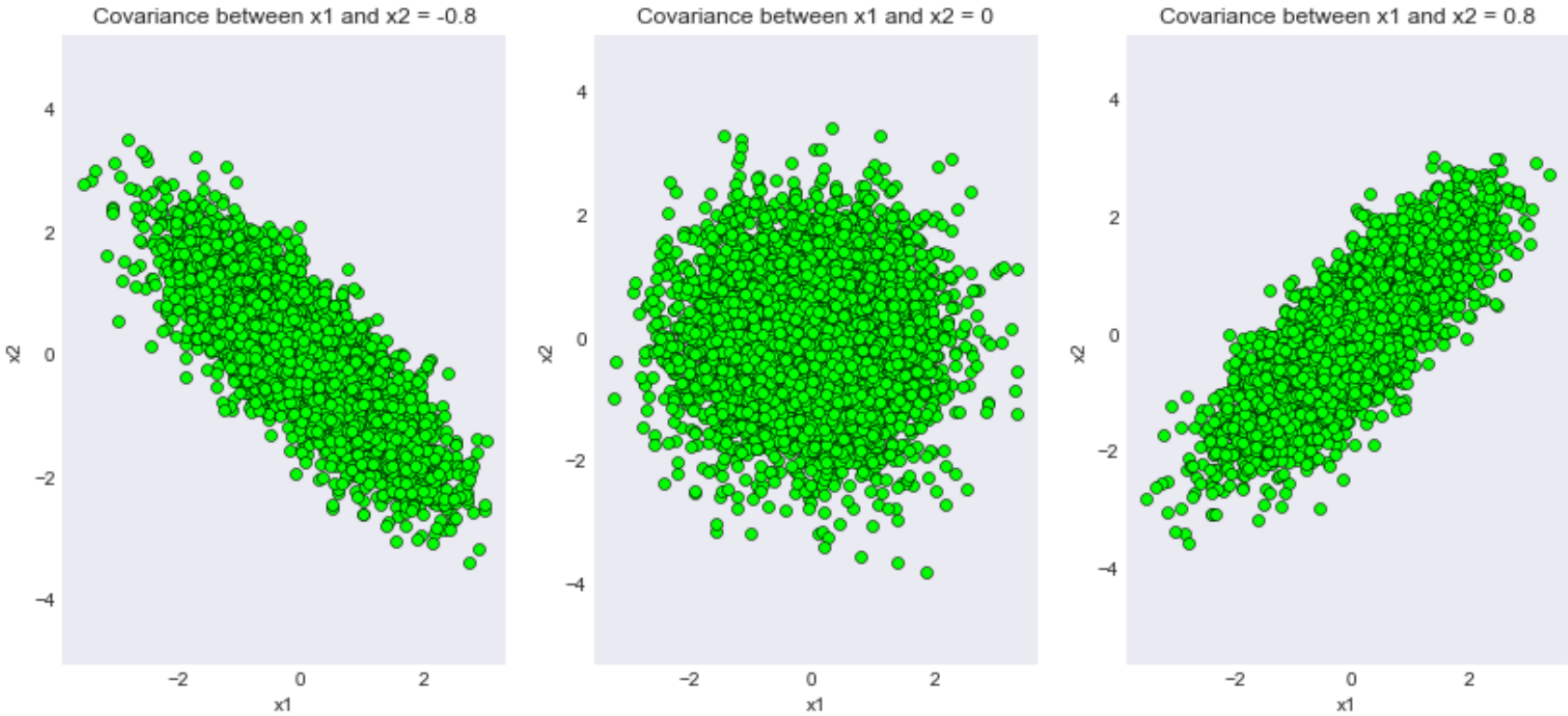
1. Si $Y \sim \mathrm{N}(\mu_{p\times 1}, \Sigma_{p\times p})$ et $B_{n\times p}$, $\theta_{n\times 1}$ sont donnés, alors
$$\theta + BY \sim \mathrm{N}(\theta + B\mu,\, B\Sigma B^\top).$$
2. Densité $\mathrm{N}(\mu, \Sigma)$, si $\Sigma \succ 0$:
$$f_Y(y) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\tfrac{1}{2}(y-\mu)^\top \Sigma^{-1}(y-\mu)\right\}.$$
3. FGM de $Y \sim \mathrm{N}(\mu, \Sigma)$:
$$M_Y(u) = \exp\left(u^\top \mu + \tfrac{1}{2} u^\top \Sigma u\right).$$
4. Si $(\mathrm{X}, \mathrm{Y})^\top = (X_1, ..., X_p, Y_1, ..., Y_q)^\top$ est gaussien,
$$\mathrm{Y} \perp\!\!\!\perp \mathrm{X} \iff \mathrm{cov}\{X_i, Y_j\} = 0 \,\forall i \in \{1, ..., p\} \;\&\; j \in \{1, ..., q\}.$$
5. Plus généralement, si $Y \sim \mathrm{N}(\mu_{p\times 1}, \Sigma_{p\times p})$, alors
$$AY \perp BY \iff A\Sigma B^\top = 0.$$
6. Les marginales et conditionelles d'une gaussienne sont aussi gaussiennes (la réciproque n'est pas vraie !).

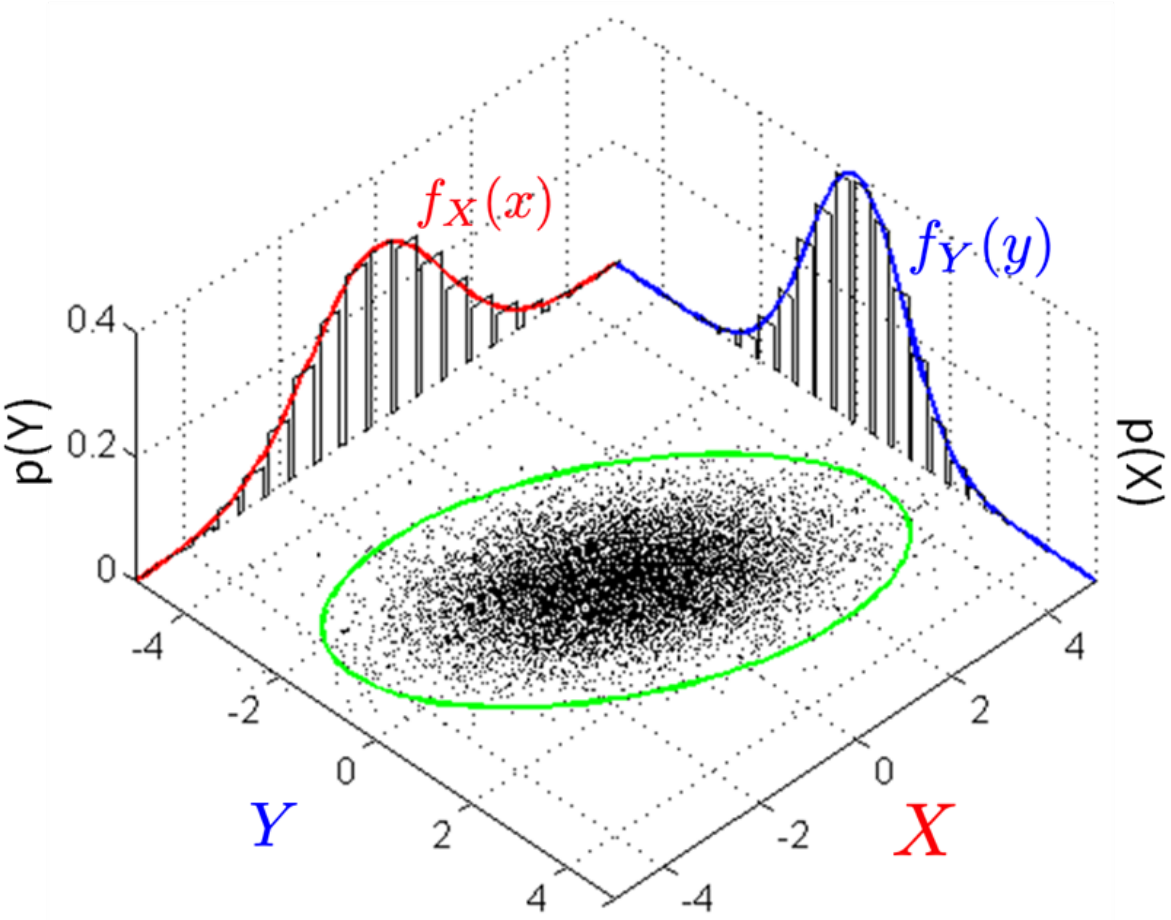

Covariance between x1 and x2 = -0.8
Covariance between x1 and x2 = 0
Covariance between x1 and x2 = 0.8
x1
x2
3
2
1
0
−1
−2
−3


Covariance between x1 and x2 = -0.8
x1
x2


Covariance between x1 and x2 = 0
x1
x2


Covariance between x1 and x2 = 0.8
x1
x2


Covariance between x1 and x2 = -0.8
Covariance between x1 and x2 = 0
Covariance between x1 and x2 = 0.8
x1
x2
4
2
0
−2
−4


$f_X(x)$
$f_Y(y)$
0.4
0.2
0
p(Y)
p(X)
-4
-2
0
2
4
Y
X

(1) On vérifie la définition de la loi gaussienne multivariée.

(2) Les variables gaussiennes centrées réduites i.i.d. ont conjointement la loi $\mathcal{N}(0, \mathrm{I})$, puis on applique une transformation de densité.

(3) On observe que $M_{\mathrm{Y}}(u) = \mathbb{E}[e^{u^\top \mathrm{Y}}] = M_{u^\top \mathrm{Y}}(1) = \exp\left(u^\top \mu + \frac{1}{2} u^\top \Sigma u\right)$.

(4) La densité se factorise si et seulement si l'on n'a pas de termes croisés dans la forme quadratique de l'exposant ; autrement dit Σ^{-1} est diagonale (ce qui équivaut à Σ diagonale).

(5) On l'obtient en considérant le cas où $M_{\binom{\mathrm{X}}{\mathrm{Y}}}\left(\binom{u}{v}\right) = M_{\mathrm{X}}(u)\, M_{\mathrm{Y}}(v)\ \forall\, u, v$.

(6) Pour les marginales : tout sous-vecteur de Y s'écrit AY pour une matrice A appropriée (par ex. $Y_j = e_j^\top \mathrm{Y}$), et l'on applique la première propriété.

Les densités conditionnelles décrivent des propriétés "bipartites" : elles impliquent des couples de vecteurs gaussiens obtenus en sélectionnant certaines coordonnées de la loi jointe. On peut vérifier cela au niveau du cas bivarié, puis l'étendre au cas général à l'aide de la notation par blocs. Nous ne faisons que le cas bivarié, qui est déjà très instructif.

Partons de la construction d'un exemple précédent avec $Z_1, Z_2 \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$ et

$$X_1 = \mu_1 + \sigma_1 Z_1,$$
$$X_2 = \mu_2 + \rho\,\sigma_2 Z_1 + \sigma_2\sqrt{1-\rho^2}\, Z_2.$$

Savoir que $X_1 = x$ fixe

$$Z_1 = \frac{x-\mu_1}{\sigma_1} \quad \text{(déterministe)},$$

tandis que $Z_2 \sim \mathrm{N}(0,1)$ reste indépendant de X_1. Ainsi

$$X_2 \,\big|\, X_1 = x = \underbrace{\mu_2 + \rho\,\sigma_2\,\frac{x-\mu_1}{\sigma_1}}_{\text{terme déterministe}} + \sigma_2\sqrt{1-\rho^2}\, Z_2,$$

c'est une image affine d'un normal standard $\Rightarrow$ $X_2|\{X_1 = x\}$ est gaussienne.
(*Si on souhaite les paramètres :* moyenne $\mu_2 + \rho\,\frac{\sigma_2}{\sigma_1}(x-\mu_1)$, variance $\sigma_2^2(1-\rho^2)$.)

Extension générale. Pour un vecteur gaussien $Y = \mu + AZ$ avec $Z \sim \mathrm{N}(0, I)$, en écrivant A par lignes et en scindant $Z = (Z_1, Z_2)$, conditionner sur un bloc linéaire fige Z_1 et laisse Z_2 gaussien indépendant $\Rightarrow$ la conditionnelle est gaussienne (argument par blocs).

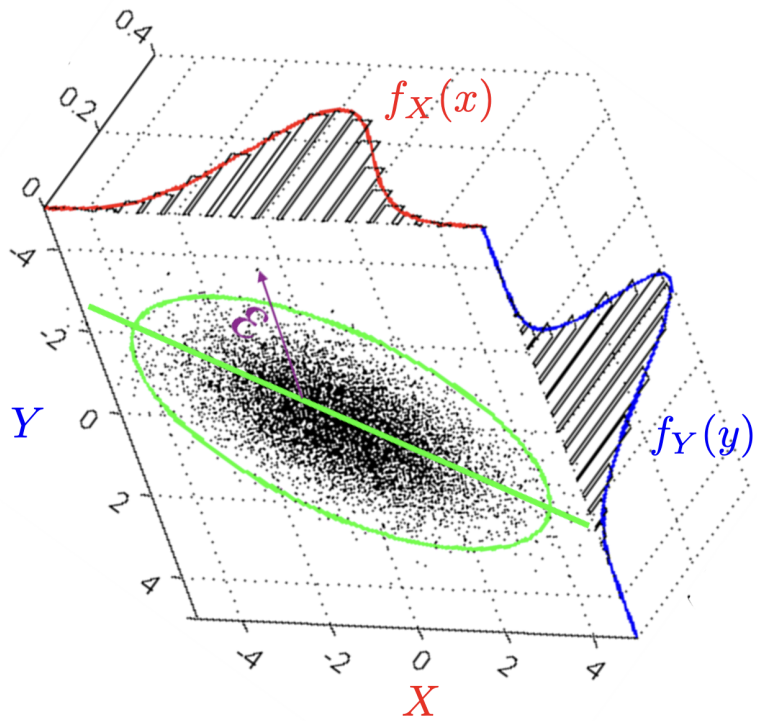
La représentation qui a mené vers la loi conditionnelle correspond à l'expression

$$Y = \beta_0 + \beta X + \varepsilon, \qquad X \sim \mathcal{N}(\mathbb{E}[X], \mathrm{var}(X)) \text{ indépendant de } \varepsilon \sim \mathcal{N}(0, \mathrm{var}(\varepsilon))$$

plus familière dans le cadre de la "régression linéaire", où :

- $\beta_0 = \mathbb{E}[X] - \frac{\mathrm{cov}\{X,Y\}}{\mathrm{var}(Y)}\mathbb{E}[Y]$ est appelé *l'ordonnée à l'origine*
- $\beta = \mathrm{cov}\{X, Y\}/\mathrm{var}(X)$ est appelé *le coefficient de régression*
- ε est appelé *l'erreur* ou *l'innovation*, qui est *homoscédastique* en ce sens que $\mathrm{var}(\varepsilon) = \mathrm{var}(Y) - \mathrm{cov}^2\{X, Y\}/\mathrm{var}(X)$ ne dépend pas de la réalisation de X.

Cela explique pourquoi on parle d'une "régression linéaire" lorsqu'on conditionne une loi normale multivariée.



Nous concluons notre discussion des lois gaussiennes en mentionnant trois lois qui leur sont liées.

Elles apparaissent lorsque l'on considère la norme au carré de vecteurs gaussiens (considérations de longueur), ou bien lorsque l'on prend des rapports de normes au carré de gaussiennes indépendantes (comparaisons de longueurs).

Les formules elles-mêmes importent moins que le fait qu'elles puissent être déterminées explicitement, et qu'elles ne dépendent d'aucun paramètre inconnu — seulement des dimensions ambiantes.

Ceci est extrêmement utile dans les problèmes d'inférence statistique (tests d'hypothèses, par exemple), comme nous le verrons par la suite.

Les trois lois fondamentales sont :

- la loi χ^2,
- la loi F de Snedecor-Fisher.
- la loi t de Student .

Proposition (Sommes de carrés gaussiens)

Soient $\{Z_1, ..., Z_k\}$ des variables aléatoires i.i.d. de loi $N(0,1)$. Alors

$$Z_1^2 + \ldots + Z_k^2 \sim \chi_k^2.$$

Une variable aléatoire X est dite suivre une loi du χ^2 de paramètre $k \in \mathbb{N}$ (appelé le nombre de degrés de liberté), notée $X \sim \chi_k^2$, si

$$f_X(x) = \begin{cases} \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-x/2}, & \text{si } x \geq 0, \\ 0, & \text{si } x < 0. \end{cases}$$

L'espérance, la variance et la fonction génératrice des moments de $X \sim \chi_k^2$ sont données par

$$\mathbb{E}[X] = k, \qquad \text{var}[X] = 2k, \qquad M(t) = (1-2t)^{-k/2}, \quad t < \tfrac{1}{2}.$$

On peut obtenir la fonction génératrice des moments en observant qu'il s'agit de la somme de k variables aléatoires indépendantes de loi χ_1^2.

Proposition (Rapports de sommes de carrés gaussiens)

Soient $Y_1 \sim \chi^2_{d_1}$ et $Y_2 \sim \chi^2_{d_2}$ deux variables aléatoires indépendantes. Alors

$$\frac{Y_1/d_1}{Y_2/d_2} \sim \mathsf{F}_{d_1,d_2}.$$

Une variable aléatoire X est dite suivre la loi de Snedecor-Fisher F de paramètres $d_1, d_2 \in \mathbb{N}$, notée $X \sim \mathsf{F}_{d_1,d_2}$, si

$$f_X(x) = \begin{cases} \frac{1}{B\left(\frac{d_1}{2},\frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{d_1/2} x^{\frac{d_1}{2}-1} \left(1+\frac{d_1}{d_2}x\right)^{-(d_1+d_2)/2}, & \text{si } x \geq 0, \\ 0, & \text{si } x < 0. \end{cases}$$

L'espérance et la variance de $X \sim \mathsf{F}_{d_1,d_2}$ sont données par

$$\mathbb{E}[X] = \frac{d_2}{d_2-2}, \text{ pour } d_2 > 2, \qquad \mathrm{var}[X] = \frac{2\,d_2^2(d_1+d_2-2)}{d_1(d_2-4)(d_2-2)^2}, \text{ pour } d_2 > 4.$$

La fonction génératrice des moments n'existe pas.

Un "cas particulier" de la loi F particulièrement important est la loi t_ν de Student

Proposition (Lien entre la loi t et la loi F)

Soient $Z \sim N(0,1)$ et $Y \sim \chi^2_\nu$ deux variables aléatoires indépendantes. Alors

$$T = \frac{Z}{\sqrt{Y/\nu}} \sim \mathrm{t}_\nu.$$

On peut réécrire T^2 comme

$$T^2 = \frac{Z^2/1}{Y/\nu}.$$

Comme $Z^2 \sim \chi^2_1$ et que $Y \sim \chi^2_\nu$ est indépendant, on a

$$T^2 \sim \mathrm{F}_{1,\nu}.$$

Ainsi, la loi t_ν peut être vue comme la racine (avec signe) d'une loi $\mathrm{F}_{1,\nu}$.

En supposant $k > 2$, l'espérance et la variance de $X \sim \mathrm{t}_k$ sont

$$\mathbb{E}[X] = 0, \qquad \mathrm{var}[X] = \frac{k}{k-2}.$$

L'espérance n'existe pas pour $k = 1$ et la variance n'existe pas pour $k \leq 2$. La fonction génératrice des moments n'existe pour aucun $k \in \mathbb{N}$.

Convergence de variables aléatoires et/ou de leurs lois

Nous voulons finalement comprendre comment une structure simple peut émerger de systèmes aléatoires complexes, "à grande échelle ou à long terme".

Cela nécessite des notions appropriées de convergence.

Nous avons déjà vu des notions de convergence pour des événements — mais il nous faut maintenant les relier aux variables aléatoires et aux vecteurs aléatoires.

Rappel : les variables aléatoires sont des *fonctions* entre des *espaces mesurables*, donc plusieurs notions de convergence peuvent être définies, par exemple :

- Convergence en loi (ou convergence faible)
- Convergence en probabilité (ou convergence en mesure)
- Convergence avec probabilité 1 (ou convergence presque sûre)
- ...

Chacune de ces notions est qualitativement différente ; nous verrons comment elles se relient les unes aux autres.

Rappelons que $X \equiv X(\omega)$ et chaque $X_n \equiv X_n(\omega)$ sont des fonctions mesurables,

$$X, X_n : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})).$$

La convergence déterministe point par point de telles fonctions signifierait que,

$$\forall\, \omega \in \Omega \;\&\; \forall \varepsilon > 0\, \exists N \geq 1 : |X_n(\omega) - X(\omega)| < \epsilon, \quad \forall n \geq N.$$

Mais nous disposons d'une mesure de probabilité sous-jacente — il faut donc tenir compte des probabilités.

Dans la plupart des cas, il suffit de pouvoir dire quelque chose de plus faible :

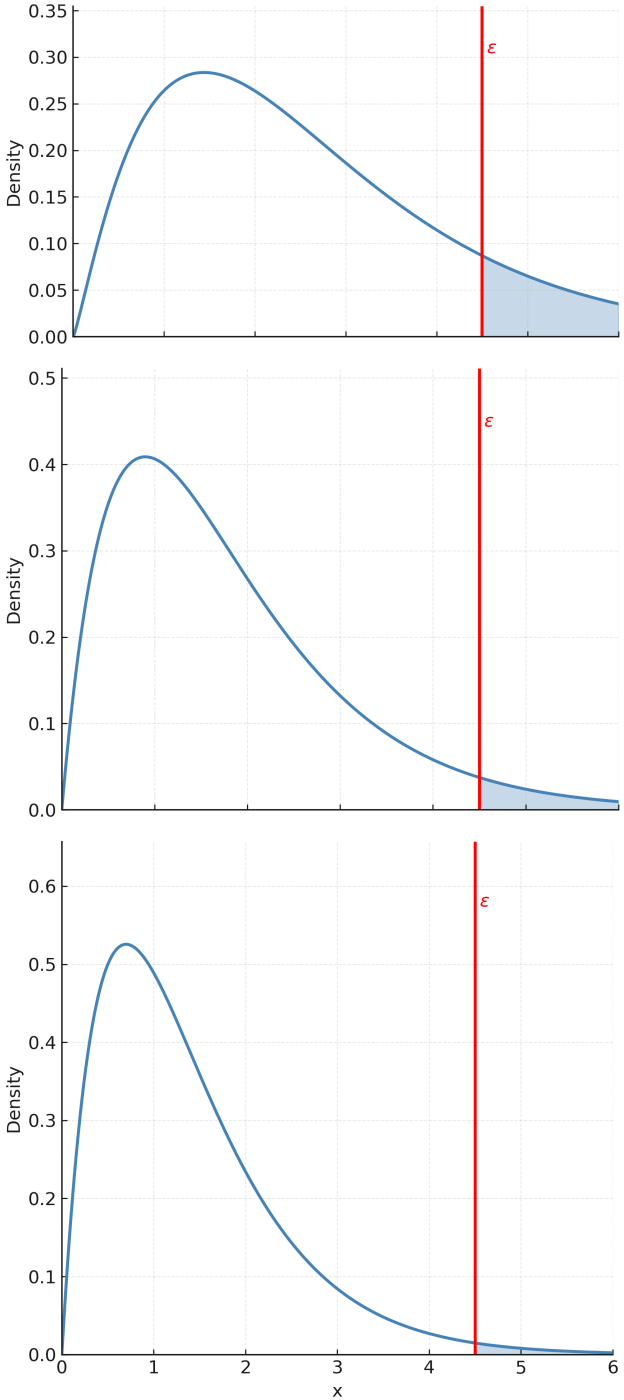
"pour n suffisamment grand, il y a une probabilité écrasante que X_n et X se réalisent très près l'un de l'autre."

Ceci est formalisé par la notion de convergence en probabilité :

Définition (Convergence en probabilité)

Soient $\{X_n\}$ et X des variables aléatoires sur $(\Omega, \mathcal{F})$. Étant donné une mesure de probabilité $\mathbb{P}$ sur $(\Omega, \mathcal{F})$, on dit que X_n converge en probabilité vers X et on écrit $X_n \xrightarrow{\mathbb{P}} X$ si

$$\forall\, \epsilon > 0, \quad \mathbb{P}[|X_n - X| > \epsilon] \overset{n\to\infty}{\longrightarrow} 0.$$



- La convergence en probabilité est une affirmation portant sur le comportement conjoint de X_n et de X.
- Pour déterminer la convergence en probabilité, il faut connaître la loi de $|X_n - X|$.
- Celle-ci peut en général être obtenue par transformation à partir de la loi conjointe $F_{(X_n,X)}$ du vecteur $(X_n, X)^\top$.
- Une fois $F_{|X_n-X|}(\cdot)$ connue, il faut examiner

$$1 - F_{|X_n-X|}(\varepsilon) = \mathbb{P}\{|X_n - X| > \varepsilon\},$$

c'est-à-dire la queue droite de la distribution.

- En pratique, il s'agit donc de montrer que $1 - F_{|X_n-X|}(\varepsilon)$ converge vers zéro lorsque $n \to \infty$, pour tout ε.
- Les graphiques à gauche montrent $1 - F_{|X_1-X|}(\varepsilon)$, $1 - F_{|X_2-X|}(\varepsilon)$, $1 - F_{|X_3-X|}(\varepsilon), \ldots$

Example

Soient $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}[0,1]$, et définissons

$$M_n = \max\{X_1, \ldots, X_n\}, \qquad \& \qquad m_n = \min\{X_1, \ldots, X_n\}.$$

Alors,

$$F_{M_n}(x) = x^n \quad \Longrightarrow \quad \mathbb{P}\big(|M_n - 1| > \varepsilon\big) = \mathbb{P}\big(M_n < 1 - \varepsilon\big) = (1-\varepsilon)^n \overset{n\to\infty}{\longrightarrow} 0,$$

pour tout $0 < \varepsilon < 1$. Ainsi, on a bien $M_n \overset{\mathbb{P}}{\longrightarrow} 1$. De façon similaire, $m_n \overset{\mathbb{P}}{\longrightarrow} 0$.
Soit $B \sim \text{Bernoulli}(1/2)$ indépendant de $\{X_j\}$, et posons

$$Y_n = B\, M_n + (1 - B)\, m_n.$$

Montrons que $Y_n \overset{\mathbb{P}}{\to} B$. Pour tout $\varepsilon \in (0,1)$,

$$\begin{aligned}
\mathbb{P}\big(|\, Y_n - B| > \varepsilon\big) &= \mathbb{P}\big(B = 1,\ |M_n - 1| > \varepsilon\big) + \mathbb{P}\big(B = 0,\ |m_n - 0| > \varepsilon\big) \\
&= \tfrac{1}{2}\, \mathbb{P}(M_n < 1 - \varepsilon) + \tfrac{1}{2}\, \mathbb{P}(m_n > \varepsilon) \\
&= \tfrac{1}{2}\,(1-\varepsilon)^n + \tfrac{1}{2}\,(1-\varepsilon)^n = (1-\varepsilon)^n \underset{n\to\infty}{\longrightarrow} 0.
\end{aligned}$$

Ce que la convergence en probabilité ne dit pas :

1. Elle ne concerne pas les trajectoires complètes $X_1(\omega), X_2(\omega), \ldots$ pour un ω donné, mais plutôt la variable $X_n(\cdot)$ (le n-ième élément, pour n grand) considérée sur tout Ω.
2. En d'autres termes, elle ne peut pas exclure qu'une trajectoire particulière $X_1(\omega), X_2(\omega), \ldots$ dépasse à plusieurs reprises une distance ϵ de $X(\omega)$.

Des énoncés comme (2) concernent non seulement l'événement $A_{n,\varepsilon} = \{|X_n - X| > \varepsilon\}$, mais l'événement "$A_{n,\varepsilon}$ se produit une infinité de fois".

En effet, comment écrit-on correctement l'événement de convergence,

$$A = \left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\} \quad ?$$

Est-il même clair que $A \in \mathcal{F}$? Heureusement, oui :

$$\begin{aligned} A &= \{\omega \in \Omega : \forall k \geq 1\, \exists N \geq 1 : |X_n(\omega) - X(\omega)| < k^{-1}\, \forall n \geq N\} \\ &= \bigcap_{k\geq 1} \bigcup_{N\geq 1} \bigcap_{n\geq N} \{\omega \in \Omega : |X_n(\omega) - X(\omega)| < 1/k\} \\ &= \bigcap_{k\geq 1} \liminf\{|X_n - X| < 1/k\} = \left(\bigcup_{k\geq 1} \limsup\{|X_n - X| \geq 1/k\}\right)^c. \end{aligned}$$

On voit donc que la probabilité que la suite converge est une question bien plus subtile. Lorsque cette probabilité vaut 1, on dit que X_n converge vers X presque sûrement :

$$X_n \xrightarrow{\text{p.s.}} X \iff \mathbb{P}\left\{\lim_{n\to\infty} |X_n - X| = 0\right\} = 1$$

$$\iff \forall \varepsilon > 0,\ \mathbb{P}\{|X_n - X| > \varepsilon \text{ i.o.}\} = 0.$$

La dernière équivalence découle de notre travail à la diapositive précédente et de la σ-sous-additivité des mesures de probabilité.

La convergence presque sûre implique la convergence en probabilité (exercice). Mais la réciproque est fausse :

Example ($\xrightarrow{\mathbb{P}}$ n'implique pas $\xrightarrow{p.s.}$)

Supposons que $Y_n \sim \text{Bernoulli}(1/n)$ soient indépendantes : chaque Y_n vaut 0 avec probabilité $1 - n^{-1}$ et 1 avec probabilité n^{-1}. Clairement, $|Y_n| \leq 1$ presque sûrement. Et, pour tout $0 < \epsilon < 1$,

$$\mathbb{P}[|Y_n| > \epsilon] = \mathbb{P}[Y_n = 1] = \tfrac{1}{n} \to 0.$$

Ainsi $Y_n \xrightarrow{\mathbb{P}} 0$. Mais la série $\sum_n \mathbb{P}[Y_n = 1]$ diverge, donc par Borel–Cantelli 2, l'événement $\{Y_n = 1\}$ se produit une infinité de fois avec probabilité 1, et par conséquent Y_n ne converge pas p.s.

Cependant, si le taux de convergence en probabilité est suffisamment rapide :

Proposition (une convergence rapide en probabilité entraîne la convergence presque sûre)

Soient $\{X_n\}$ et X des variables aléatoires sur le même espace de probabilité. Si, pour tout $\varepsilon > 0$, la probabilité $\mathbb{P}(|X_n - X| \geq \varepsilon)$ décroît suffisamment vite en n pour que

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) < \infty,$$

alors $X_n \to X$ presque sûrement.

Démonstration.

Fixons $k \in \mathbb{N}$ et posons $A_{n,k} = \{|X_n - X| \geq 1/k\}$. Par hypothèse avec $\varepsilon = 1/k$, $\sum_{n=1}^{\infty} \mathbb{P}(A_{n,k}) < \infty$. D'après le premier lemme de Borel–Cantelli,

$$\mathbb{P}\big(\limsup_{n\to\infty} A_{n,k}\big) = 0 \qquad \text{pour tout } k \in \mathbb{N}.$$

En utilisant la sous-additivité et en sommant sur k,

$$\mathbb{P}\Big(\bigcup_{k=1}^{\infty} \limsup_{n\to\infty} A_{n,k}\Big) \leq \sum_{k=1}^{\infty} \mathbb{P}\big(\limsup_{n\to\infty} A_{n,k}\big) = 0.$$

□

- Jusqu'à présent, nous avons essentiellement examiné l'évolution de $F_{|X_n - X|}(\cdot)$, pour tous les arguments.
- Que se passe-t-il si, au lieu de cela, nous regardons l'évolution de $|F_{X_n}(\cdot) - F_X(\cdot)|$ pour tous les arguments ?
- Cela compare les lois marginales de X_n et de X — et non leur comportement conjoint.

Convergence en loi (ou convergence faible)

Soit $\{F_n\}_{n\geq 1}$ une suite de fonctions de répartition, et F une fonction de répartition sur $\mathbb{R}$. On dit que F_n converge en loi vers F, et on écrit $F_n \xrightarrow{d} F$, lorsque

$$F_n(x) \overset{n\to\infty}{\longrightarrow} F(x), \quad \forall\, x \notin D_F$$

(convergence point-par-point sauf aux points de discontinuité de la limite F).

Remarque : Nous avons formulé cette forme de convergence uniquement en termes de fonctions de répartition — afin de souligner qu'il s'agit, par nature, d'une propriété portant sur les distributions elles-mêmes.
Bien sûr, dans la plupart des cas, $F_n(x)$ et $F(x)$ proviennent des lois de variables aléatoires X_n et X ; on écrit alors $X_n \xrightarrow{d} X \iff F_{X_n} \overset{n\to\infty}{\longrightarrow} F_X$.

Example (Le maximum de variables aléatoires uniformes)

Soient $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}(0,1)$, $M_n = \max\{X_1, \ldots, X_n\}$, et $Q_n = n(1 - M_n)$.

$$\mathbb{P}[Q_n \leq x] = \mathbb{P}[M_n \geq 1 - x/n] = 1 - \left(1 - \frac{x}{n}\right)^n \overset{n\to\infty}{\longrightarrow} 1 - e^{-x}.$$

Notez que la limite est la fonction de répartition d'une variable aléatoire $\text{Exp}(1)$.

Établir la convergence en loi est plus simple que la convergence en probabilité: pas nécessaire de manipuler la loi jointe, il suffit de manipuler les lois marginales.

Conformément à cette idée de plus simple, on peut en fait montrer que (exercise):

$$X_n \overset{\mathbb{P}}{\to} X \implies X_n \overset{d}{\to} X$$

On obtient ainsi la chaîne suivante,

$$X_n \overset{p.s.}{\to} X \implies X_n \overset{\mathbb{P}}{\to} X \implies X_n \overset{d}{\to} X$$

ce qui explique le terme de convergence faible pour " $\overset{d}{\to}$ "
(la plus faible que nous ayons rencontrée).

Lemme de Slutsky

Soient (X_n) et (Y_n) deux suites de variables aléatoires telles que

$$X_n \xrightarrow{d} X \qquad \text{et} \qquad Y_n \xrightarrow{\mathbb{P}} c \in \mathbb{R}.$$

Alors :

(a) $X_n + Y_n \xrightarrow{d} X + c$;

(b) $X_n Y_n \xrightarrow{d} cX$.

- Attention : nous ne pouvons pas remplacer la constante déterministe c par une variable aléatoire (exercice: trouver un contre-exemple).

Preuve.

(a) On peut supposer $c = 0$. Soit x un point de continuité de F_X. On a :

$$\begin{aligned} \mathbb{P}[X_n + Y_n \leq x] &= \mathbb{P}[X_n + Y_n \leq x, \, |Y_n| \leq \varepsilon] + \mathbb{P}[X_n + Y_n \leq x, \, |Y_n| > \varepsilon] \\ &\leq \mathbb{P}[X_n \leq x + \varepsilon] + \mathbb{P}[|Y_n| > \varepsilon]. \end{aligned}$$

De même,

$$\mathbb{P}[X_n \leq x - \varepsilon] \;\leq\; \mathbb{P}[X_n + Y_n \leq x] + \mathbb{P}[|Y_n| > \varepsilon].$$

Ainsi,

$$\mathbb{P}[X_n \leq x-\varepsilon] - \mathbb{P}[|Y_n| > \varepsilon] \;\leq\; \mathbb{P}[X_n + Y_n \leq x] \;\leq\; \mathbb{P}[X_n \leq x+\varepsilon] + \mathbb{P}[|Y_n| > \varepsilon].$$

Comme ε est arbitraire, on obtient le résultat de (a) en faisant tendre $n \to \infty$.

(b) Il suffit encore de supposer $c = 0$ (car $(Y_n + c)X_n = X_n Y_n + cX_n$, et la partie (a) donnera la conclusion). Soient $\varepsilon, M > 0$. Alors :

$$\begin{aligned}
\mathbb{P}[|X_n Y_n| > \varepsilon] \quad &\leq \quad \mathbb{P}[|X_n Y_n| > \varepsilon,\ |Y_n| \leq 1/M] + \mathbb{P}[|Y_n| \geq 1/M] \\
&\leq \quad \mathbb{P}[|X_n| > \varepsilon M] + \mathbb{P}[|Y_n| \geq 1/M] \\
&\xrightarrow[n\to\infty]{} \quad \mathbb{P}[|X| > \varepsilon M] + 0.
\end{aligned}$$

Le premier terme peut être rendu arbitrairement petit en faisant tendre $M \to \infty$. D'où la conclusion. □

Remarquons que pour toute variable aléatoire Y, on peut toujours définir

$$1_{\{Y \leq x\}} \sim \text{Bernoulli}(F(x)), \quad \text{où } F(x) = \mathbb{E}\left[1_{\{Y \leq x\}}\right].$$

Ainsi, on peut reformuler la convergence en loi $X_n \xrightarrow{d} X$ comme la condition

$$\mathbb{E}\left[1_{\{X_n \leq x\}}\right] \overset{n\to\infty}{\longrightarrow} \mathbb{E}\left[1_{\{X \leq x\}}\right], \quad \forall\, x \notin D_F.$$

On peut approximer des fonctions en escalier par des "marches lissées", alors :

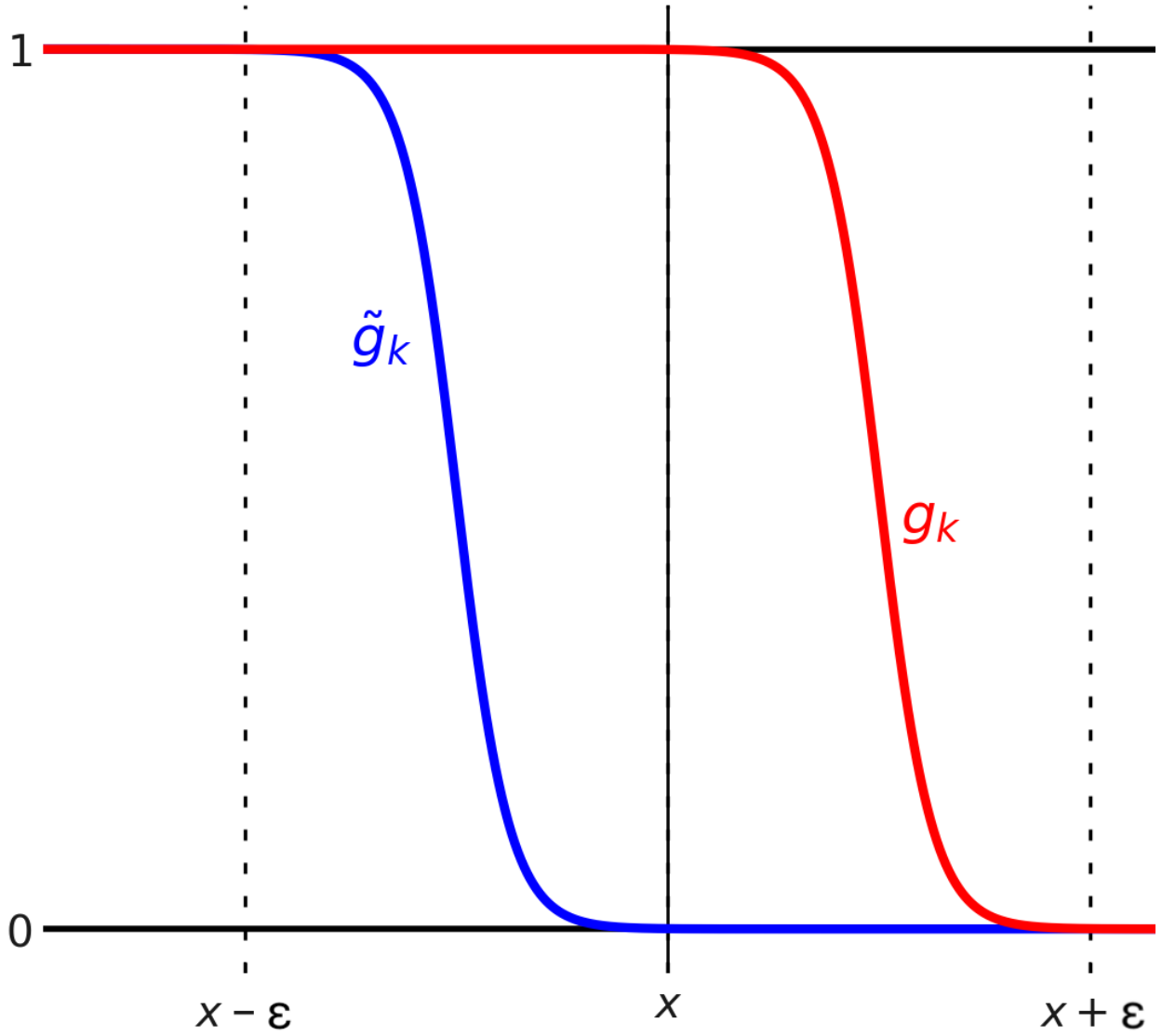
Lemme (Portmanteau)

Étant données des variables aléatoires X_n de loi F_{X_n} et X de loi F, on a $X_n \xrightarrow{d} X$ si et seulement si

$$\mathbb{E}[g(X_n)] \overset{n\to\infty}{\longrightarrow} \mathbb{E}[g(X)],$$

pour toute fonction bornée $g(\cdot)$ de classe C^k, avec un certain $k \geq 0$.

Remarquons que la convergence des moments ne suffit pas en général — c'est-à-dire que, dans l'énoncé du théorème, prendre (même tous) les monômes de la forme $g(x) = x^k$ ne permet pas de conclure.


1
0
$\tilde{g}_k$
g_k
$x - \varepsilon$
x
$x + \varepsilon$

Démonstration (suffisance)

Nous ne prouverons que la partie "si", qui est tout ce dont nous avons besoin dans le cours. Considérons un point de continuité $x \notin D_F$, et prenons $\varepsilon > 0$. Soient $\tilde{g}, g : \mathbb{R} \to [0,1]$ des interpolantes décroissantes de classe C^k reliant respectivement les points $\{(x-\varepsilon, 1), (x, 0)\}$ et $\{(x, 1), (x+\varepsilon, 0)\}$. Il s'ensuit que

$$1_{\{u \leq x-\varepsilon\}} \leq \tilde{g}(u) \leq 1_{\{u \leq x\}} \leq g(u) \leq 1_{\{u \leq x+\varepsilon\}}$$

D'après la troisième inégalité, l'hypothèse du lemme $(*)$, puis la dernière inégalité,

$$\limsup_{n\to\infty} F_{X_n}(x) \leq \lim_{n\to\infty} \mathbb{E}[g(X_n)] \overset{*}{=} \mathbb{E}[g(X)] \leq \mathbb{E}\big[1_{\{X \leq x+\varepsilon\}}\big] = F_X(x+\varepsilon)$$

D'après la deuxième inégalité, l'hypothèse du lemme $(*)$, puis la première inégalité,

$$\liminf_{n\to\infty} F_{X_n}(x) \geq \lim_{n\to\infty} \mathbb{E}[\tilde{g}(X_n)] \overset{*}{=} \mathbb{E}[\tilde{g}(X)] \geq \mathbb{E}\big[1_{\{X \leq x-\varepsilon\}}\big] = F_X(x-\varepsilon)$$

En combinant les deux bornes, on obtient

$$F_X(x-\varepsilon) \leq \liminf_{n\to\infty} F_{X_n}(x) \leq \limsup_{n\to\infty} F_{X_n}(x) \leq F_X(x+\varepsilon).$$

Le résultat s'en déduit puisque F est continue en x et que ε est arbitraire. □

La loi des grands nombres et le théorème central limite

Imaginons un système aléatoire désordonné, constitué d'une grande collection de variables aléatoires indépendantes et identiquement distribuées,

$$X_1, X_2, X_3, \ldots, X_n \qquad (n \text{ grand}).$$

Que se passe-t-il si l'on applique à ce système une fonction scalaire déterministe ?

$$(X_1, \ldots, X_n) \mapsto g(X_1, \ldots, X_n)$$

Transformer le système modifie sa loi — nous avons vu comment la calculer.

- Mais les composantes du système n'agissent pas en symphonie, plutôt dans une cacophonie aléatoire.
- On pourrait s'attendre à ce que leurs contributions s'annulent mutuellement...
- ... à condition que chaque composante contribue à la fonction de manière à peu près égale, et de plus en plus négligeable lorsque n croît.

Si tel est le cas, on peut s'attendre à ce que la sortie de g appliqué au système présente un comportement très ordonné — presque constant — malgré la nature très chaotique de son entrée.

Peut-être la fonction la plus simple de ce type est

$$g(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

c'est-à-dire l'état moyen du système.

- Chaque composante contribue de manière égale à la fonction.
- La contribution de chaque composante s'annule asymptotiquement...
- ... à condition que la probabilité d'être arbitrairement grande soit petite.

Ces considérations conduisent au premier grand théorème limite de la théorie des probabilités :

Théorème (Loi forte des grands nombres)

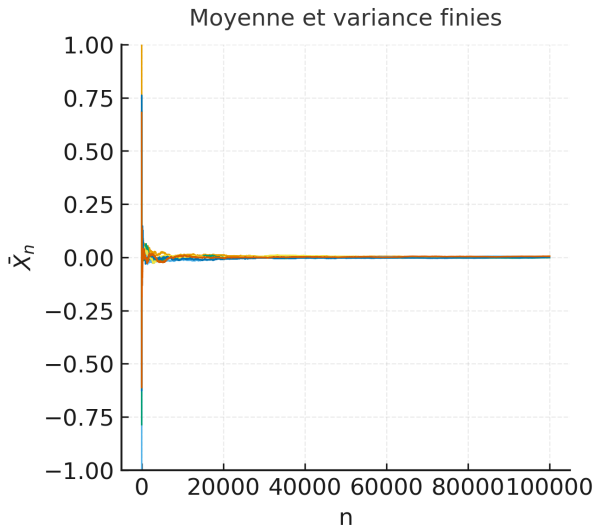
Soit $\{X_i\}_{i \geq 1}$ une suite de v.a. i.i.d. vérifiant $\mathbb{E}|X_i| < \infty$. Alors,

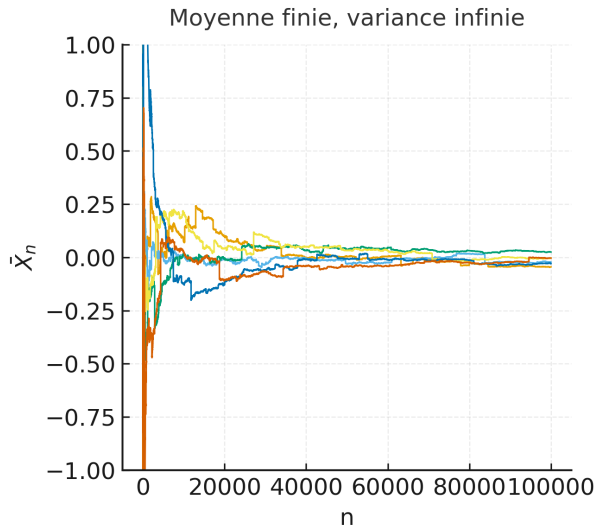
$$\bar{X}_n \equiv \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{\text{p.s.}} \mathbb{E}[X_i].$$

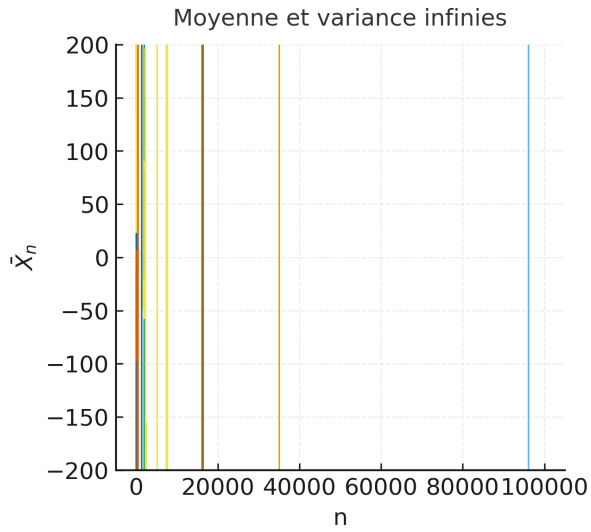
Réfléchissez à ce que cela signifie pour des variables de Bernoulli...

... nous pouvons enfin légitimer l'interprétation de la probabilité comme une limite de fréquences relatives — un résultat, non une définition !

Comportement de la moyenne empirique selon les moments de X_i







Démonstration (sous hypothèse de moments d'ordre 4)

On donne une preuve sous l'hypothèse plus forte que $m_4 \equiv \mathbb{E}[X_i^4] < \infty$. On suppose $\mathbb{E}[X_i] = 0$ et $\mathrm{Var}[X_i] = 1$ (sinon on travaille avec les versions centrées et réduites). On peut développer $\mathbb{E}\big[\bar{X}_n^4\big]$ en regroupant les indices :

$$\frac{1}{n^4}\,\mathbb{E}\Bigg[\sum_i X_i^4 + 4\sum_{i\neq j} X_i X_j^3 + 3\sum_{i\neq j} X_i^2 X_j^2 + 6\sum_{i,j,k} X_i X_j X_k^2 + \sum_{i\neq j\neq k\neq \ell} X_i X_j X_k X_\ell\Bigg]$$

Tous les termes contenant au moins une variable à la première puissance ont une espérance nulle (indépendance et $\mathbb{E}[X_i] = 0$). Il reste donc

$$\mathbb{E}[\bar{X}_n^4] = \frac{1}{n^4}\Big(n\,\mathbb{E}[X_1^4] + 3n(n-1)\,\mathbb{E}[X_1^2 X_2^2]\Big) \leq \frac{m_4}{n^3} + \frac{3}{n^2}.$$

Par l'inégalité de Markov : $\mathbb{P}\big(|\bar{X}_n| > \varepsilon\big) \leq \dfrac{\mathbb{E}[\bar{X}_n^4]}{\varepsilon^4} \leq \dfrac{1}{\varepsilon^4}\left(\dfrac{m_4}{n^3} + \dfrac{3}{n^2}\right).$

Pour tout $\varepsilon > 0$, cette borne est sommable en n, donc Borel–Cantelli donne

$$\forall \varepsilon > 0, \quad \mathbb{P}\big(|\bar{X}_n| > \varepsilon \text{ i.o.}\big) = 0.$$

□

Aussi simple que semble la loi forte, elle a des conséquences surprenantes.

Supposons que $g : \mathbb{R}^n \to \mathbb{R}$ soit continûment différentiable et que

$$\left| \frac{\partial g}{\partial x_i}(x) \right| \leq \frac{C}{n} \quad \text{pour tout } i = 1, \ldots, n.$$

Alors :

- La fonction est également sensible à chaque coordonnée — g dépend du système *collectivement*, et non d'une variable isolée. C'est la symétrie.
- La sensibilité totale, mesurée par $\sum_{i=1}^n \left| \frac{\partial g}{\partial x_i} \right|$, reste bornée lorsque n croît. C'est la lissité : g ne peut pas osciller brutalement.

Avec ces notions de symétrie et de lissité, la loi des grands nombres implique que

“Une fonction lisse et symétrique de nombreuses variables aléatoires i.i.d. d'espérance finie est presque sûrement constante.”

Corollaire de la loi forte des grands nombres

Soit $\{X_i\}_{i\geq 1}$ une suite de v.a. i.i.d. vérifiant $\mathbb{E}[|X_1|] < \infty$, et soit $\mu = \mathbb{E}[X_1]$. Si $g_n : \mathbb{R}^n \to \mathbb{R}$ est une suite de fonctions lisses et symétriques à n arguments, au sens défini précédemment, alors

$$g_n(X_1, \ldots, X_n) - g_n(\mu, \ldots, \mu) \xrightarrow{\text{p.s.}} 0 \quad \text{quand } n \to \infty.$$

Démonstration.

Par le théorème des accroissements finis (développement de Taylor d'ordre 0),

$$g_n(X_1, \ldots, X_n) = g_n(\mu, \ldots, \mu) + \sum_{i=1}^{n} \frac{\partial g_n}{\partial x_i}(\xi_n)\,(X_i - \mu),$$

pour un certain vecteur ξ_n entre $(X_1, \ldots, X_n)$ et $(\mu, \ldots, \mu)$. Donc

$$|g_n(X_1, \ldots, X_n) - g_n(\mu, \ldots, \mu)| \leq \frac{C}{n} \sum_{i=1}^{n} |X_i - \mu| \xrightarrow{\text{p.s.}} 0$$

par la loi forte des grands nombres. □

Supposons maintenant une variance finie. Soient $X_1, ..., X_n$ des v.a. i.i.d. de moyenne μ et de variance σ^2. Alors, avec $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$,

$$\mathbb{E}[\bar{X}_n] = \mu \qquad \& \qquad \text{var}(\bar{X}_n) = \sigma^2/n.$$

L'échelle en $1/n$ tue la variance asymptotiquement, de sorte que la loi de $\bar{X}_n$ se contracte vers sa moyenne :

$$\mathbb{P}[|\bar{X}_n - \mu| > \varepsilon] \leq \frac{\sigma^2}{n\varepsilon^2} \overset{n\to\infty}{\longrightarrow} 0 \quad \text{[Tchebychev]}$$

Pour apprécier les fluctuations de $\bar{X}_n$ autour de μ, il faut "zoomer" :

- Quelle est la bonne mise à l'échelle ?
- Pour garder la variance vivante prenons $\sqrt{n}$:

$$\text{var}\{\sqrt{n}(\bar{X}_n - \mu)\} = n\frac{\sigma^2}{n} = \sigma^2, \qquad \forall\, n \geq 1.$$

Nous avons donc la quantité cible : la standardisation de la moyenne empirique,

$$\sqrt{n}(\bar{X}_n - \mu)/\sigma$$

Quelle devrait être une limite distributionnelle candidate ?

Les indices :

- Nous introduisons de plus en plus d'aléa indépendant (entropie), tout en maintenant les premier et deuxième moments constants.
- Les lois gaussiennes standards sont stables par opération de moyennage/standardisation :

$$\text{si } X_i \overset{iid}{\sim} \mathrm{N}(0,1), \text{ alors } \sqrt{n}(\bar{X}_n - \mu) \sim \mathrm{N}(0,1), \quad \forall\, n \geq 1.$$

Ces deux éléments nous conduisent à conjecturer :

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

Et c'est bien le cas :

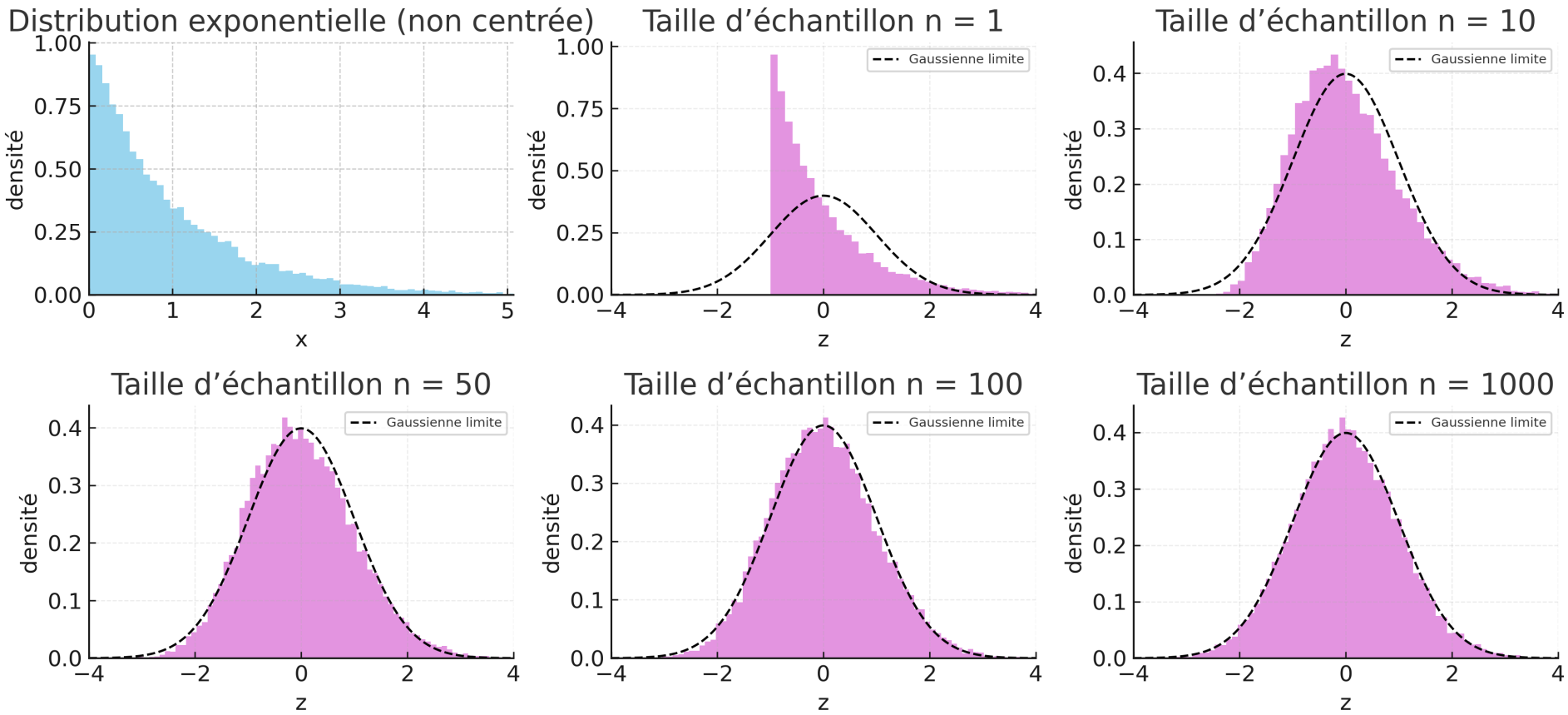
Théorème (Théorème Central Limite)

Soit $\{X_i\}_{i\geq 1}$ une suite de variables aléatoires i.i.d. de moyenne μ et de variance finie $\sigma^2 < \infty$. Alors,

$$\sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} \mathrm{N}(0,1), \quad \text{quand } n \to \infty.$$

Instance la plus élémentaire de universalité : les choix microscopiques de modélisation importent peu à grande échelle – seuls comptent les grands traits (les deux premiers moments ici).

TCL: moyennes d'un échantillon exponentiel (asymétrie prononcée)



Le TCL est utilisé pour approcher des probabilités impliquant des sommes de variables aléatoires indépendantes. Sous les conditions précédentes, on a

$$\mathbb{E}\left(\sum_{j=1}^{n} X_j\right) = n\mu, \qquad \mathrm{Var}\left(\sum_{j=1}^{n} X_j\right) = n\sigma^2,$$

donc

$$\frac{\sum_{j=1}^{n} X_j - n\mu}{\sqrt{n\sigma^2}} = \frac{n(\bar{X} - \mu)}{\sqrt{n\sigma^2}} = \frac{n^{1/2}(\bar{X} - \mu)}{\sigma} = Z_n$$

peut être approximé par une variable normale :

$$\mathbb{P}\left(\sum_{j=1}^{n} X_j \leq x\right) = \mathbb{P}\left\{\frac{\sum_{j=1}^{n} X_j - n\mu}{\sqrt{n\sigma^2}} \leq \frac{x - n\mu}{(n\sigma^2)^{1/2}}\right\} \approx \Phi\left(\frac{x - n\mu}{(n\sigma^2)^{1/2}}\right).$$

Preuve (sous la condition supplémentaire que $\mathbb{E}|X_i|^3 < \infty$)

Supposons sans perte de généralité que $\mu = 0$ et $\sigma^2 = 1$. Établir

$$\mathbb{E}\left[g\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right)\right] \xrightarrow{d} \mathbb{E}[g(Z_0)] \equiv \mathbb{E}\left[g\left(\frac{Z_1 + \cdots + Z_n}{\sqrt{n}}\right)\right], \quad \forall\, g \in C_b^3(\mathbb{R}),$$

où $\{Z_k\}_{k\geq 0} \overset{iid}{\sim} \mathrm{N}(0,1)$ nous permettra de conclure par le lemme portmanteau.
Soit $\Delta_i = X_i - Z_i$ et $\Delta = (\Delta_1, \ldots, \Delta_n)^\top = X - Z$. Définissons

$$\begin{aligned} s(u_1, ..., u_n) &= \frac{1}{\sqrt{n}}(u_1 + \ldots + u_n) \quad G_n(u_1, ..., u_n) = g(s(u)) \\ h(t) &= G_n(Z + t\Delta), \quad t \in [0,1]. \end{aligned}$$

où $h(t)$ est le chemin d'interpolation entre $h(0) = G_n(\mathrm{Z})$ et $h(1) = G_n(\mathrm{X})$.
Par développement de Taylor du second ordre de $h(t)$ en 0, avec reste intégral,

$$h(t) - h(0) = h'(0)t + \frac{1}{2}h''(0)t^2 + \tfrac{1}{2}\int_0^t (t-y)^2 h^{(3)}(y)\, dy.$$

$$\implies G_n(X) - G_n(Z) = h(1) - h(0) = h'(0) + \tfrac{1}{2}h''(0) + \tfrac{1}{2}\int_0^1 (1-t)^2 h^{(3)}(t)\, dt.$$

Comme $\mathbb{E}X_i = \mathbb{E}Z_i = 0$ et $\mathbb{E}X_i^2 = \mathbb{E}Z_i^2 = 1$: $\mathbb{E}[h'(0)] = 0$, $\mathbb{E}[h''(0)] = 0$ (exercice : vérifiez-le). Et, en remarquant que toutes les dérivées partielles de G_n coïncident,

$$h^{(3)}(t) = \frac{g^{(3)}(s(Z+t\Delta))}{n^{3/2}} \Big(\sum_{i=1}^{n} \Delta_i \Big)^3, \quad s(u) = \tfrac{u_1 + \cdots + u_n}{\sqrt{n}}.$$

Ainsi

$$\big| \mathbb{E}G_n(X) - \mathbb{E}G_n(Z) \big| \leq \frac{\|g^{(3)}\|_\infty}{3!\, n^{3/2}} \, \big| \mathbb{E} \big(\sum_{i=1}^{n} \Delta_i \big)^3 \big|.$$

En développant le cube et en utilisant l'indépendance et la centration,

$$\mathbb{E} \left(\sum_{i=1}^{n} \Delta_i \right)^3 = \sum_{i,j,k=1}^{n} \mathbb{E}[\Delta_i \Delta_j \Delta_k] = \sum_{i=j=k=1}^{n} \mathbb{E}[\Delta_i \Delta_j \Delta_k] = \sum_{i=1}^{n} \mathbb{E}[\Delta_i^3].$$

(tous les termes mixtes s'annulent car au moins un facteur est de moyenne nulle).

$$\implies \big| \mathbb{E}g(\sqrt{n}\bar{X}_n) - \mathbb{E}g(\sqrt{n}\bar{Z}_n) \big| \leq \frac{\|g^{(3)}\|_\infty}{3!\sqrt{n}} \, \mathbb{E}|X_1 - Z_1|^3 \propto \frac{\sum_{p+q=3} \mathbb{E}|X_1|^p \mathbb{E}|Z_1|^q}{\sqrt{n}}$$

Parfois, on s'intéresse plutôt à une fonction de la somme qu'à la somme elle-même — que faire ? Devons-nous transformer la limite, ou bien... ?

- Remarquons que si $g : \mathbb{R} \to \mathbb{R}$ est une fonction régulière, alors $g_n(X_1, \ldots, X_n) \equiv g(\overline{X}_n)$ est "lisse et symétrique" au sens vu précédemment.
- On peut donc espérer que le TCL s'applique directement à $g(\overline{X}_n)$, sans qu'il soit nécessaire de transformer quoi que ce soit ?

La méthode Delta affirme précisément cela :

Théorème (Méthode Delta)

Soient $X_1, X_2, \ldots$ des variables aléatoires indépendantes d'espérance μ et de variance $0 < \sigma^2 < \infty$, et soit $g'(\mu) \neq 0$, où g' est la dérivée de g. Alors

$$\frac{g(\bar{X}_n) - g(\mu)}{\{g'(\mu)^2 \sigma^2 / n\}^{1/2}} \xrightarrow{D} \mathrm{N}(0,1), \qquad n \to \infty.$$

En effet, cela signifie que pour n grand, on peut utiliser l'approximation

$$g(\overline{X}_n) \overset{d}{\approx} \mathrm{N}\big(g(\mu), \sigma^2[g'(\mu)]^2\big).$$

Des résultats analogues peuvent être établis pour des fonctions lisses/symétriques plus générales de $\{X_1, \ldots, X_n\}$, mais nous nous arrêterons ici.

Définition (Convergence en loi vectorielle)

Soit $\{X_n\}$ une suite de vecteurs aléatoires de $\mathbb{R}^d$, et X un vecteur aléatoire de $\mathbb{R}^d$. Soient $F_n, F : \mathbb{R}^d \to [0,1]$ les fonctions de répartition jointes correspondantes. On dit que $X_n \xrightarrow{d} X$ si

$$F_{X_n}(x) \overset{n\to\infty}{\longrightarrow} F_X(x)$$

pour tout point de continuité $x \in \mathbb{R}^d$ de F_X.

Il existe un lien entre la convergence faible scalaire et vectorielle :

Théorème (Dispositif de Cramér–Wold)

Soit $\{X_n\}$ une suite de vecteurs aléatoires de $\mathbb{R}^d$, et X un vecteur aléatoire de $\mathbb{R}^d$. Alors,

$$X_n \xrightarrow{d} X \Leftrightarrow \theta^\top X_n \xrightarrow{d} \theta^\top X, \ \forall \theta \in \mathbb{R}^d.$$

Exercice : montrer par contre-exemple que la convergence faible séparée de chaque coordonnée n'implique pas la convergence faible du vecteur aléatoire.

Nous ne démontrerons pas le dispositif de Cramér–Wold, mais nous ferons simplement quelques commentaires :

- Remarquez que la définition de la convergence faible dit essentiellement que

$$\mathbb{E}\left[1_{\{X_n \in A\}}\right] \overset{n\to\infty}{\longrightarrow} \mathbb{E}\left[1_{\{X \in A\}}\right], \quad \forall A \in \mathcal{A},$$

  où $\mathcal{A}$ est la collection de tous les quadrants tels que F_X soit continu sur leur frontière.

- Le dispositif de Cramér–Wold indique que l'on peut remplacer la classe $\mathcal{A}$ par une autre classe d'ensembles déterminante pour la convergence, essentiellement celle de tous les demi-espaces (puisqu'on peut aussi prendre des décalages déterministes).

- Connaître la loi de $\theta^\top X$ pour tout $\theta \in \mathbb{R}^d$ détermine la loi de X – car les opérations dénombrables sur les demi-espaces engendrent la σ-algèbre borélienne (elles engendrent tous les rectangles).

- Mais cela concerne une seule loi, et non une suite de lois. La version séquentielle peut sembler facile mais elle est, en réalité, non triviale, même si l'intuition de base reste correcte.

Nous pouvons maintenant donner les analogues vectoriels de LGN/TCL/Δ :

Loi des grands nombres

Soit $\{X_n\}$ une suite i.i.d. de vecteurs aléatoires telle que $\mathbb{E}\|X_1\| < \infty$, et $\mathbb{E}X_1 = \mu$. Alors,

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{a.s.} \mu$$

La démonstration est la même — coordonnée par coordonnée !

Théorème Central Limite

Soit $\{X_m\}$ une suite i.i.d. de vecteurs aléatoires dans $\mathbb{R}^d$ de moyenne μ et de covariance Σ avec $\text{trace}\{\Sigma\} < \infty$. Soit $\bar{X}_n := \sum_{i=1}^{n} X_i/n$. Alors,

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} Z \sim N(0, \Sigma).$$

Exercice : prouvez ce TCL en utilisant le TCL scalaire et Cramér–Wold.

Méthode Delta

Dans le même contexte que le théorème précédent, si $g(\cdot) : \mathbb{R}^p \to \mathbb{R}^q$ a une dérivée non nulle en μ, alors $\sqrt{n}(g(\bar{X}_n) - g(\mu)) \xrightarrow{d} [(\nabla g)(\mu)]Z$.

De la probabilité à la statistique

Probabilités :

1. Le phénomène d'intérêt est conceptualisé comme un modèle probabiliste
2. On utilise le modèle pour apprendre la probabilité des résultats possibles.
3. La probabilité est le langage de la modélisation scientifique

Statistique :

1. Le phénomène d'intérêt est conceptualisé comme un modèle probabiliste
2. On utilise des données pour apprendre quelque chose sur le modèle.
3. La statistique est le langage de l'expérimentation scientifique

"Apprendre quelque chose" peut signifier :

- Quel membre d'une famille de modèles a engendré les données ?
- Quel ensemble de modèles est compatible avec un jeu de données donné ?
- Les données sont-elles plus cohérentes avec un modèle qu'avec un autre ?

Les trois problèmes statistiques que nous allons considérer sont:

1. **Estimation**. Etant donné un échantillon $X_1, \ldots, X_n$ tiré d'une distribution F_θ qui dépend d'un paramètre inconnu θ, comment peut-on construire un estimateur, i.e une fonction de l'échantillon, dont le but est d'estimer θ?

2. **Intervalles de confiance**. Plutôt que de tenter d'estimer la valeur précise du paramètre θ qui a généré notre échantillon $X_1, \ldots, X_n$, est-ce qu'on peut construire un ensemble de valeurs sous la forme d'un intervalle, qui aura une grande probabilité de contenir le vrai paramètre θ?

3. **Tests d'hypothèses**. Etant donnée une valeur plausible θ_0 pour θ (ou plusieurs valeurs plausibles formant un ensemble Θ_0), est-ce que, sur la base de l'échantillon $X_1, \ldots, X_n$, cette valeur (ou cet ensemble) est un bon indicateur de la vraie valeur de θ?

Une notion clé qui joue un rôle important dans les trois problèmes est celle de la

vraisemblance.

La statistique comme "probabilité inverse". Considerons le cas discret.

Point de vue Probabilités

Si on se dispose d'un paramètre $\theta \in \Theta$, alors pour $(x_1, ..., x_n) \in \mathcal{X}^n$, on peut évaluer

$$(x_1, ..., x_n) \mapsto \mathbb{P}_\theta[X_1 = x_1, ..., X_n = x_n]$$

c'est à dire, comment se varie la probabilité comme fonction de l'échantillon (=du résultat).

Point de vue Statistiques

Si on se dispose d'un échantillon $(x_1, ..., x_n) \in \mathcal{X}^n$, alors pour tout $\theta \in \Theta$ on peut évaluer

$$\theta \mapsto \mathbb{P}_\theta[X_1 = x_1, ..., X_n = x_n]$$

c'est à dire, comment se varie la probabilité comme fonction du paramètre (=du modéle).

Intuition : on imagine que les θ plausibles à partir du connaissance de l'échantillon sont ceux qui rendent notre échantillon assez probable...

Definition (La vraisemblance pour une collection discrète iid)

Soit $X_1, \ldots, X_n$ une collection de variables aléatoires discrétes, indépendantes et identiquement distribuées de fonction de masse $f(x;\theta)$, où $\theta \in \mathbb{R}^p$. La vraisemblance de θ est définie par

$$L : \Theta \to [0,1]$$

$$L(\theta) = \prod_{i=1}^{n} f(X_i;\theta).$$

Remarques :

1. La vraisemblance est une fonction aléatoire
2. La vraisemblance est, en effet, $\prod_{i=1}^{n} f(X_i;\theta)$ vue comme fonction de θ
3. La vraisemblance n'est pas "la probabilité de θ"
4. La vraisemblance $L(\theta)$ est la réponse à la question : *quelle est la probabilité de l'échantillon observé lorsque le paramètre est égal à θ*

Et le cas continu? On utilisera la même définition, avec la densité au lieu de la fonction de masse, même si l'inteprération change un peu :

Definition (La vraisemblance pour une collection continue iid)

Soit $X_1, \ldots, X_n$ une collection de variables aléatoires continues, indépendantes et identiquement distribuées de fonction de densité $f(x;\theta)$, où $\theta \in \mathbb{R}^p$. La vraisemblance de θ est définie par

$$L : \Theta \to [0, +\infty)$$

$$L(\theta) = \prod_{i=1}^{n} f(X_i;\theta).$$

Remarques :

1. Notons que maintenant la vraisemblance prend de valeurs dans $\mathbb{R}_+$ entier.
2. Puisque $F(x+\epsilon/2\,;\theta) - F(x-\epsilon/2;\theta) \approx \epsilon f(x;\theta)$ lorsque $\epsilon \downarrow 0$, nous pouvons voir $\epsilon^n L(\theta)$ comme étant la probabilité approximative d'un echantillon "proche" a ce que nous avons observé.

Avant d'entrer dans le vif du sujet, un mot de terminologie. Que nous estimions, donnions des bornes de confiance ou testions une hypothèse...

Tout ce que nous ferons sera une fonction $T(X_1, \ldots, X_n)$ de l'échantillon

Cela motive la notion suivante :

Définition (statistique)

Une *statistique* est toute fonction mesurable T dont le domaine est l'espace d'échantillons $\mathcal{X}^n$, mais qui ne dépend d'aucun paramètre inconnu.

$\hookrightarrow$ Intuitivement, toute fonction qui peut être évaluée uniquement à partir de l'échantillon, et pour laquelle on peut formuler des énoncés probabilistes.
$\hookrightarrow$ Quand on applique T sur l'échantillon on obtient une variable aléatoire, avec sa propre loi de probabilité. Cette loi est appellée la loi d'echantillonage.

Example

$n^{-1} \sum_{i=1}^n X_i$ et $\max\{X_1, \ldots, X_n\}$ sont de statistiques (n est connu !).

Pour alléger la notation, il est d'usage d'écrire simplement T au lieu de $T(X_1, \ldots, X_n)$ — tout comme on écrit souvent X au lieu de $X(\omega)$ pour une v.a.

Estimation ponctuelle

Etant donné un échantillon $X_1, \ldots, X_n$ tiré d'une distribution F_θ qui dépend d'un paramètre inconnu $\theta \in \Theta$, comment peut-on estimer θ ?

Définition (Estimateur ponctuel)

Une statistique prenant des valeurs dans l'espace paramètre Θ est appelée un *estimateur ponctuel*. Réciproquement, un estimateur ponctuel est une statistique

$$T : \mathcal{X}^n \to \Theta.$$

- Puisque l'objectif d'un estimateur est de fournir une estimation du vrai θ qui a généré les données, nous le dénotons typiquement $\hat{\theta} \equiv \hat{\theta}(X_1, ..., X_n)$.
- Donc, un estimateur est une variable aléatoire. Une realisation d'un estimateur est une estimation.
- Une procédure générale que l'on peut appliquer à n'importe quel modèle concret pour produire un estimateur est appelée une méthode d'estimation.

Dans ce qui suit, on supposera que le modèle est identifiable :

$$\theta_1 \neq \theta_2 \implies F_{\theta_1} \neq F_{\theta_2}.$$

Sinon, le problème d'estimation n'est pas bien posé.

Soit $X_1, \ldots, X_n$ une collection de variables aléatoires indépendantes et identiquement distribuées de fonction de densité/masse $f(x;\theta)$, où $\theta \in \mathbb{R}^p$. La vraisemblance de θ est définie par

$$L(\theta) = \prod_{i=1}^{n} f(X_i;\theta).$$

Rapellons l'intuition : on imagine que les θ plausibles à partir du connaissance de l'échantillon sont ceux qui rendent notre échantillon assez probable...

Définition (Estimateur du maximum de vraisemblance)

Soit $X_1, \ldots, X_n$ un échantillon aléatoire iid tiré d'une distribution F_θ de fonction de densité/masse $f(x;\theta)$ et soit $\hat{\theta}$ tel que

$$L(\theta) \leq L(\hat{\theta}), \qquad \forall\, \theta \in \Theta.$$

Alors $\hat{\theta}$ est appelé un *estimateur du maximum de vraisemblance* (EMV) de θ.

- Notons que l'EMV est défini indirectement, comme l'optimum d'une fonction objective. Alors comment le déterminer?
- Lorsque la vraisemblance est une fonction dérivable de θ, le maximum de la fonction $L(\theta)$ doit être une solution de l'équation

$$\nabla_\theta L(\theta) = 0,$$

- Avant de déclarer qu'une solution $\hat{\theta}$ de cette équation est un EMV, nous devons d'abord vérifier que c'est bien un maximum (et non un minimum!).
- Si la vraisemblance est deux fois dérivable, ceci peut être fait en vérifiant que

$$-\nabla^2_\theta L(\theta)\big|_{\theta=\hat{\theta}} \succ 0,$$

  i.e que (-1) multiplié par la matrice hessienne est définie positive.
- Lorsque le paramètre est de dimension un, ceci se réduit à vérifier que la seconde dérivée est négative lorsqu'elle est évaluée à la solution de l'équation de vraisemblance.

- Afin de résoudre $\nabla_\theta L(\theta) = 0$, il faut déterminer la dérivée d'un produit de n fonctions, ce qui peut être un calcul fastidieux.
- Afin d'éviter ceci, nous nous concentrons habituellement à maximiser la *log-vraisemblance*

$$\ell(\theta) := \log L(\theta)$$

  au lieu de la vraisemblance.
- Puisque la fonction log est monotone, la vraisemblance et la log-vraisemblance ont les maximums et les minimums pour les mêmes θ.
- L'avantage de la log-vraisemblance est que nous travaillons avec une somme de n fonctions plutôt qu'un produit,

$$\ell(\theta) = \log\left(\prod_{i=1}^{n} f(X_i;\theta)\right) = \sum_{i=1}^{n} \log f(X_i;\theta).$$

- Encore une fois, si la fonction log-vraisemblance est deux fois dérivable, un EMV $\hat{\theta}$ de θ satisfera

$$\nabla_\theta \ell(\theta)\big|_{\theta=\hat{\theta}} = 0 \quad \& \quad -\nabla^2_\theta \ell(\theta)\big|_{\theta=\hat{\theta}} \succ 0.$$

Example (EMV pour la loi de Bernoulli)

Soit $X_1, \ldots, X_n \overset{iid}{\sim} Bern(p)$ et supposons que nous voulons utiliser la méthode du maximum de vraisemblance afin de construire un estimateur de $p \in (0,1)$. La vraisemblance est :

$$L(p) = \prod_{i=1}^{n} f(X_i; p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^{n} X_i}(1-p)^{n-\sum_{i=1}^{n} X_i}.$$

En prenant le logarithme de chaque côté de l'équation, nous obtenons la fonction de log-vraisemblance

$$\ell(p) = \log p \sum_{i=1}^{n} X_i + \log(1-p)\left(n - \sum_{i=1}^{n} X_i\right).$$

Nous pouvons noter que cette fonction est deux fois dérivable par rapport à p et calculer

$$\frac{d}{dp}\ell(p) = p^{-1}\sum_{i=1}^{n} X_i - (1-p)^{-1}\left(n - \sum_{i=1}^{n} X_i\right).$$

Résoudre l'équation $\ell'(p) = 0$ en fonction de p est équivalent à résoudre

$$p^{-1} \sum_{i=1}^{n} X_i - (1-p)^{-1} \left(n - \sum_{i=1}^{n} X_i \right) = 0,$$

et nous pouvons voir que cette dernière équation à un unique racine donnée par $\frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}$. Appelons cette racine $\hat{p}$, nous devons maintenant vérifier qu'elle correspond bien à un maximum. Notez que

$$\frac{d^2}{dp^2} \ell(p) = -p^2 \sum_{i=1}^{n} X_i - (1-p)^{-2} \left(n - \sum_{i=1}^{n} X_i \right),$$

et que cette expression est toujours non-positive, car $0 \leq \sum_{i=1}^{n} X_i \leq n$ presque sûrement et $p \in (0,1)$. Ainsi

$$\hat{p} = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

est l'unique EMV de p. □

Example (EMV pour la loi exponentielle)

Soit $X_1, \ldots, X_n \overset{iid}{\sim} Exp(\lambda)$ et supposons que nous voulons utiliser la méthode du maximum de vraisemblance afin de construire un estimateur de $\lambda \in (0, \infty)$. La vraisemblance est :

$$L(\lambda) = \prod_{i=1}^{n} f(X_i; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n \exp\left\{ -\lambda \sum_{i=1}^{n} X_i \right\}.$$

En prenant le logarithme de chaque côté de l'équation, nous obtenons la fonction de log-vraisemblance

$$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} X_i.$$

Nous pouvons noter que cette fonction est deux fois dérivable par rapport à λ et calculer

$$\frac{d}{d\lambda} \ell(\lambda) = n\lambda^{-1} - \sum_{i=1}^{n} X_i.$$

Résoudre l'équation $\ell'(\lambda) = 0$ en fonction de λ nous donne l'unique racine

$$\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right)^{-1} = 1/\overline{X}.$$

Appelons celle-ci $\hat{\lambda}$, nous devons maintenant vérifier qu'elle correspond bien à un maximum. Notez que

$$\frac{d^2}{d\lambda^2} \ell(\lambda) = -\frac{n}{\lambda^2}$$

et que cette expression est toujours négative, car $\lambda > 0$. Ainsi

$$\hat{\lambda} = \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right)^{-1} = 1/\overline{X}$$

est l'unique EMV de λ. □

Example (EMV pour la loi gaussienne)

Soit $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ et supposons que nous voulons utiliser la méthode du maximum de vraisemblance afin de construire un estimateur de $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$. La vraisemblance est :

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(X_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left\{ -\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{2\sigma^2} \right\}.$$

En prenant le logarithme de chaque côté de l'équation,

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2.$$

Noter que les dérivés secondes par rapport à μ et σ^2 existent et

$$\begin{aligned} \frac{\partial}{\partial \mu} \ell(\mu, \sigma^2) &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) \\ \frac{\partial}{\partial \sigma^2} \ell(\mu, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (X_i - \mu)^2. \end{aligned}$$

Résoudre l'équation $\nabla_{(\mu,\sigma^2)}\ell(\mu,\sigma^2)=0$ en fonction de (μ,σ^2) donne un système de deux équations à deux inconnues. L'unique solution de ce système est

$$\left(\overline{X}, n^{-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right).$$

Appelons cette solution $(\hat{\mu},\hat{\sigma}^2)$, nous devons maintenant vérifier qu'elle correspond bien à un maximum. Notez que

$$\frac{\partial^2}{\partial\mu^2}\ell(\mu,\sigma^2)=-\frac{n}{\sigma^2},\quad \frac{\partial^2}{\partial(\sigma^2)^2}\ell(\mu,\sigma^2)=\frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{i=1}^{n}(X_i-\mu)^2$$

$$\frac{\partial^2}{\partial\mu\partial\sigma^2}\ell(\mu,\sigma^2)=\frac{\partial^2}{\partial\sigma^2\partial\mu}\ell(\mu,\sigma^2)=-\frac{\sum_{i=1}^{n}(X_i-\mu)}{\sigma^4}=\frac{n\mu-n\overline{X}}{\sigma^4}.$$

En évaluant ces dérivés secondes en $(\hat{\mu},\hat{\sigma}^2)$, nous obtenons

$$\left.\frac{\partial^2}{\partial\mu^2}\ell(\mu,\sigma^2)\right|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)}=-\frac{n}{\hat{\sigma}^2},\quad \left.\frac{\partial^2}{\partial(\sigma^2)^2}\ell(\mu,\sigma^2)\right|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)}=-\frac{n}{2\hat{\sigma}^4}$$

$$\frac{\partial^2}{\partial\mu\partial\sigma^2}\ell(\mu,\sigma^2)\Big|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)} = \frac{\partial^2}{\partial\sigma^2\partial\mu}\ell(\mu,\sigma^2)\Big|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)} = \frac{n\hat{\mu}-n\hat{\mu}}{\hat{\sigma}^4} = 0.$$

Nous obtenons que la matrice

$$\left[-\nabla^2_{(\mu,\sigma^2)}\ell(\mu,\sigma^2)\Big|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)}\right]$$

est diagonale. Afin de montrer qu'elle est définie positive, il suffit de montrer que les éléments de sa diagonale sont positifs. C'est bien le cas ici, puisque $\hat{\sigma}^2$ est positif avec probabilité 1. Ainsi l'unique EMV de (μ,σ^2) est donné par

$$(\hat{\mu},\hat{\sigma}^2) = \left(\overline{X}\ ,\ \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right).$$

□

Il apparaît clairement que les dérivées aléatoires $\ell'(\cdot)$ et $-\ell''(\cdot)$ jouent un rôle fondamental dans la détermination de l'EMV d'un $\theta_0 \in \Theta \subseteq \mathbb{R}$.

Que peut-on dire de leur "champ moyen" (espérance) $\mathbb{E}[\ell'(\cdot)]$ et $\mathbb{E}[-\ell''(\cdot)]$?

Lemme (Identités de Bartlett)

Si le modèle est identifiable et la vraisemblance est suffisamment régulière pour permettre sa double dérivation sous le signe intégral, alors (exercice) :

1. $\boxed{\mathbb{E}_\theta[\ell'(\theta)] = 0}$
2. $\boxed{\mathbb{E}_\theta[-\ell''(\theta)] = \mathbb{E}_\theta\left\{[\ell'(\theta)]^2\right\} = \mathrm{var}_\theta[\ell'(\theta)] \in (0, \infty).}$

- En moyenne, la dérivée s'annule au vrai paramètre.
- En moyenne, la seconde dérivée est négative au vrai paramètre.
- Lorsqu'on evalue $\ell'(\cdot)$ au vrai θ, elle devient une variable aléatoire de moyenne zero et variance finie et égale à la courbure de $\ell'(\cdot)$ au vrai θ.
- Comme $\ell'(\cdot)$ est continue, on a aussi $\mathrm{var}_\theta\{\ell'(\theta \pm \varepsilon)\} < \infty$ pour $\epsilon > 0$ petit.

Alors quand l'échange dérivée/intégrale est permise, Bartlett dit :

"le vrai paramètre est un maximum de la log-vraisemblance moyenne"

Comme la log-vraisemblance (fois $1/n$) convergera vers sa moyenne (LGN), est-ce qu'on peut ésperer que son maximum convergera vers le vrai paramètre?

En une dimension nous pouvons énoncer un résultat assez simple :

Théorème (Convergence en probabilité de l'EMV)

En plus des conditions dans le lemme de Bartlett, supposons que l'EMV $\hat{\theta}_n$ existe de manière unique pour tout n. Alors, si θ_0 est le vrai paramètre,

$$\mathbb{P}_{\theta_0}\{|\hat{\theta}_n - \theta_0| > \epsilon\} \overset{n\to\infty}{\longrightarrow} 0$$

Un estimateur qui converge en probabilité vers le vrai paramètre quand la taille n de l'echantillon diverge est dit cohérent.

Preuve.

Développons la moyenne de $\ell'(\cdot)$ autour de θ_0 (définition de la dérivée de ℓ' en θ, et échange dérivée–espérance justifié par la supposition de régularité) :

$$\mathbb{E}_{\theta_0}\big[\ell'(\theta)\big] = \mathbb{E}_{\theta_0}\big[\ell'(\theta_0)\big] + (\theta - \theta_0)\,\mathbb{E}_{\theta_0}\big[\ell''(\theta_0)\big] + o\big(|\theta - \theta_0|\big).$$

Les deux identités de Bartlett permettent de conclure

$$\mathbb{E}_{\theta_0}\big[\ell'(\theta)\big] = -(\theta - \theta_0)\,c + o\big(|\theta - \theta_0|\big).$$

avec $c = \mathrm{var}_{\theta_0}\{\ell'(\theta_0)\} \in (0, \infty)$. Donc, pour $\varepsilon > 0$ petit et $\theta = \theta_0 \pm \varepsilon$,

$$\delta_- \;:=\; \mathbb{E}_{\theta_0}\big[\ell'(\theta_0 - \varepsilon)\big] \;>\; 0, \qquad \delta_+ \;:=\; \mathbb{E}_{\theta_0}\big[\ell'(\theta_0 + \varepsilon)\big] \;<\; 0.$$

Autrement dit, le $\ell'(\cdot)$ moyen est positif à gauche et négatif à droite de θ_0.

Comme $\mathrm{var}_{\theta_0}\{\ell'(\theta_0)\} < \infty$ et $\ell'(\cdot)$ est continue, on a aussi $\mathrm{var}_{\theta_0}\{\ell'(\theta_0 \pm \varepsilon)\} < \infty$ pour tout $\epsilon > 0$ assez petit, et

$$\ell'(\theta_0 \pm \varepsilon) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i, \theta_0 \pm \varepsilon) \xrightarrow{\text{p.s.}} \mathbb{E}_{\theta_0}\big[\ell'(\theta_0 \pm \varepsilon)\big] = \delta_\pm, \qquad \text{[LGN]}$$

donc $\mathbb{P}\big[\ell'(\theta_0 - \varepsilon) > 0\big] \to 1$ et $\mathbb{P}\big[\ell'(\theta_0 + \varepsilon) < 0\big] \to 1$.

Alors par inclusion/exclusion,

$$\mathbb{P}\big[\,\ell'(\theta_0-\varepsilon)>0 \quad \& \quad \ell'(\theta_0+\varepsilon)<0\,\big]\to 1.$$

Maintenant, par définition de l'EMV,

$$\ell'(\hat{\theta}_n)=0, \quad \text{et ce zéro est unique (par supposition).}$$

Comme f est C^1, la fonction $\theta\mapsto\ell'(\theta)$ est continue. Ainsi, si pour $\varepsilon>0$ on a

$$\ell'(\theta_0-\varepsilon)<0 \quad \text{et} \quad \ell'(\theta_0+\varepsilon)>0,$$

alors, par le théorème des valeurs intermédiaires: $\exists\,\theta\in(\theta_0-\varepsilon,\theta_0+\varepsilon):\ \ell'(\theta)=0$. Par l'unicité du zéro il s'ensuit que $\hat{\theta}_n\in(\theta_0-\varepsilon,\theta_0+\varepsilon)$. En autre terme,

$$\{\ell'(\theta_0-\varepsilon)<0\ \&\ \ell'(\theta_0+\varepsilon)>0\}\subseteq\{\theta_0-\varepsilon<\hat{\theta}_n<\theta_0+\varepsilon\},$$

et par monotonicité de probabilité,

$$\underbrace{\mathbb{P}\big[\ell'(\theta_0-\varepsilon)<0,\ \ell'(\theta_0+\varepsilon)>0\big]}_{\to 1}\le\mathbb{P}\big[\theta_0-\varepsilon<\hat{\theta}_n<\theta_0+\varepsilon\big]\le\mathbb{P}\{|\hat{\theta}_n-\theta_0|\le\epsilon\}.$$

□

Théorème (TCL pour la EMV)

En plus des conditions supposées dans le dernier Théorème, supposons que l'espace des paramètres $\Theta \subseteq \mathbb{R}$ soit ouvert, et que la log-vraisemblance admette une troisième dérivée "intégrablement bornée" proche à θ_0 (voir ci-dessous). Alors, si θ_0 est le vrai paramètre,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathrm{N}\left(0\,,\ \frac{1}{I(\theta_0)}\right).$$

où $I(\theta) = \mathbb{E}\left[-\partial_\theta^2 \log f(X_1;\theta)\right] = \mathbb{E}\left[\left(\partial_\theta f(X_1;\theta)\right)^2\right]$ est l'information de Fisher.

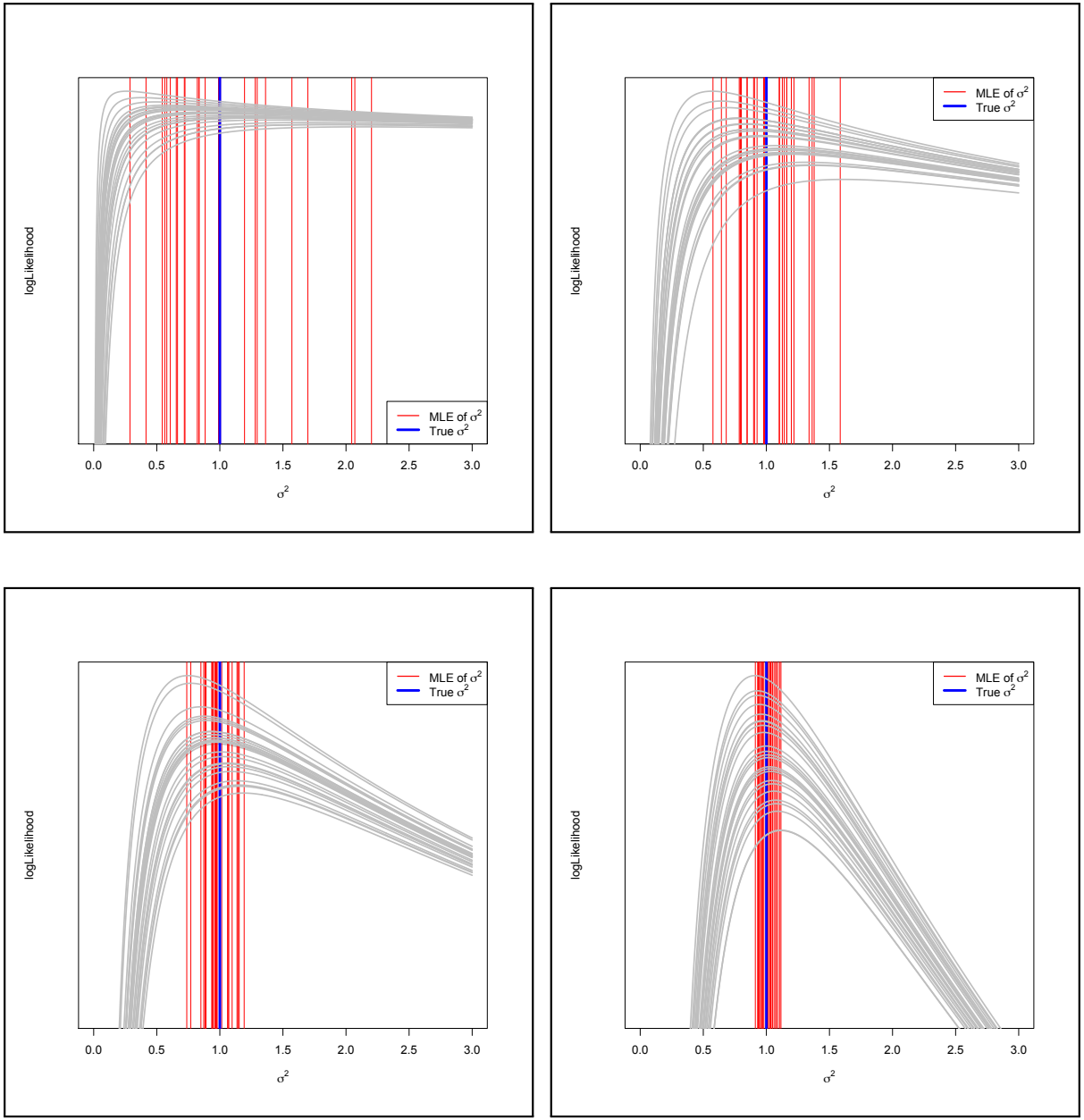
- On peut interpréter ceci comme

$$\hat{\theta}_n \overset{d}{\approx} \mathrm{N}\left(\theta\,,\ \frac{1}{nI(\theta)}\right), \qquad \text{pour } n \text{ grand.}$$

- La condition selon laquelle ℓ''' est "intégrablement bornée proche à θ_0" exige que : $\exists\ M_0(x) > 0$ et $\delta_0 > 0$ tels que $\mathbb{E}_{\theta_0}[M_0(X_i)] < \infty$ et

$$|\theta - \theta_0| < \delta_0 \implies |\ell'''(x;\theta)| \leq M_0(x).$$

Pourquoi $nI(\theta)$? (... courbure) Prenons $n = 10, 50, 150, 450$.



Preuve.

Soit $\ell_i(\theta) = \log f(X_i;\theta)$, et alors $\ell(\theta) = \sum_{i=1}^n \ell_i(\theta)$ et le termes $\ell_i(\theta)$ sont i.i.d. Si $\hat{\theta}_n$ est l'unique maximum de la vraisemblance, on a

$$\ell'(\hat{\theta}_n) \equiv \sum_{i=1}^n \ell_i'(\hat{\theta}_n) = 0.$$

En développant cette équation en série de Taylor proche de θ_0, on obtient

$$0 = \sum_{i=1}^n \ell_i'(\hat{\theta}_n) = \sum_{i=1}^n \ell_i'(\theta_0) + (\hat{\theta}_n - \theta_0)\sum_{i=1}^n \ell_i''(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 \sum_{i=1}^n \ell_i'''(\theta_n^*),$$

où θ_n^* est compris entre θ_0 et $\hat{\theta}_n$. En divisant par $\sqrt{n}$, on obtient :

$$0 = \frac{1}{\sqrt{n}}\sum_{i=1}^n \ell_i'(\theta) + \sqrt{n}(\hat{\theta}_n - \theta_0)\frac{1}{n}\sum_{i=1}^n \ell_i''(\theta_0) + \frac{1}{2}\sqrt{n}(\hat{\theta}_n - \theta_0)^2 \frac{1}{n}\sum_{i=1}^n \ell_i'''(\theta_n^*).$$

$$\implies \sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{n^{-1/2}\sum_{i=1}^n \ell_i'(\theta_0)}{n^{-1}\sum_{i=1}^n -\ell_i''(\theta_0) - (\hat{\theta}_n - \theta_0)(2n)^{-1}\sum_{i=1}^n \ell_i'''(\theta_n^*)}.$$

D'après le TCL et le lemme de Bartlett, il vient que

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell_i'(\theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)).$$

Ensuite, la loi LGN, combinée au lemme de Bartlett, implique

$$\frac{1}{n} \sum_{i=1}^{n} -\ell_i''(\theta_0) \xrightarrow{p} I(\theta_0).$$

Par le lemme de Slutsky, le théorème suivra si l'on montre que $R_n := (\hat{\theta}_n - \theta_0) \frac{1}{2n} \sum_{i=1}^{n} \ell_i'''(\theta_n^*) \xrightarrow{\mathbb{P}} 0$. Cela est établi dans le lemme suivant. □

Lemme

Dans le même contexte que le théorème précédent,

$$R_n := (\hat{\theta}_n - \theta_0) \frac{1}{2n} \sum_{i=1}^{n} \ell_i'''(\theta_n^*) \xrightarrow{\mathbb{P}} 0$$

pour toute variable aléatoire θ_n^* située sur le segment reliant $\hat{\theta}_n$ et θ.

Preuve.

Pour tout $\epsilon > 0$, on a

$$\mathbb{P}[|R_n| > \epsilon] = \underbrace{\mathbb{P}[|R_n| > \epsilon, |\hat{\theta}_n - \theta_0| > \delta]}_{\leq \mathbb{P}[|\hat{\theta}_n - \theta_0| > \delta] \to 0} + \mathbb{P}[|R_n| > \epsilon, |\hat{\theta}_n - \theta_0| \leq \delta].$$

Si $|\hat{\theta}_n - \theta_0| < \delta$, la borne intégrable donne $|R_n| \leq \frac{\delta}{2n} \sum_{i=1}^{n} M(X_i) = \bar{M}_n$. Ainsi,

$$\mathbb{P}[|R_n| > \epsilon, |\hat{\theta}_n - \theta| \leq \delta] \leq \mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta \bar{M}_n].$$

Pour $\xi > 0$, ce dernier terme se borne par

$$\mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta \bar{M}_n, \bar{M}_n \leq M + \xi] + \mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta \bar{M}_n, \bar{M}_n > M + \xi]$$

qui à son tour est majoré par

$$\mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta(M + \xi)] + \mathbb{P}[|\bar{M}_n - M| > \xi].$$

Mais la loi des grands nombres implique que

$$\bar{M}_n = \frac{1}{n} \sum_{i=1}^{n} M(X_i) \overset{\text{p.s.}}{\to} \mathbb{E}[M(X_1)] < \infty.$$

Il en résulte que

$$\mathbb{P}[|\bar{M}_n - M| > \xi] \to 0.$$

Comme on peut toujours choisir δ aussi petit que souhaité*, on peut rendre la limite du terme

$$\mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta(M + \xi)]$$

nul. En résumé, on a établi que $R_n \xrightarrow{p} 0$. □

* On a $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ si et seulement s'il existe une suite $(a_n)_{n\geq 1}$ décroissante, $a_n \downarrow 0$, telle que

$$\mathbb{P}\Big(|\hat{\theta}_n - \theta| > a_n\Big) < a_n \qquad (n \geq 1).$$

($\Leftarrow$) Soit $\varepsilon > 0$. Pour n assez grand, $a_n \leq \varepsilon$, donc

$$\mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) \leq \mathbb{P}(|\hat{\theta}_n - \theta| > a_n) < a_n \to 0.$$

($\Rightarrow$) Posons $\varepsilon_k = 2^{-k}$. Comme $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$, pour tout k il existe N_k tel que, pour tout $n \geq N_k$,

$$\mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon_k) < \varepsilon_k.$$

Définissons $a_n := \varepsilon_k$ si $N_k \leq n < N_{k+1}$ (en prenant (N_k) croissante). Alors $a_n \downarrow 0$ et, pour $N_k \leq n < N_{k+1}$,

$$\mathbb{P}(|\hat{\theta}_n - \theta| > a_n) = \mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon_k) < \varepsilon_k = a_n.$$

Alors, pour être precis, nous replacons δ par a_n dans la preuve.

On étend la vraisemblance au cas $\theta \in \Theta \subseteq \mathbb{R}^d$. Les conditions de Bartlett deviennent :

1. $\boxed{\mathbb{E}_\theta[\nabla \ell(\theta)] = 0}$
2. $\boxed{\mathbb{E}_\theta[-\nabla^2 \ell(\theta)] = \mathbb{E}_\theta\left\{[\nabla \ell(\theta)]^2\right\} = \mathrm{var}_\theta[\nabla \ell(\theta)] \succ 0 \ (\prec \infty).}$

La cohérence est plus délicate à établir, mais tient sous des conditions quasi minimales pour les familles exponentielles. On obtient :

$$\sqrt{n}\,(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathrm{N}\big(0,\, I(\theta_0)^{-1}\big)\,,$$

où

$$I(\theta) = \mathbb{E}\big[-\nabla_\theta^2 \log f(X_1;\theta)\big] = \mathbb{E}\big[\nabla_\theta \log f(X_1;\theta)\, \nabla_\theta \log f(X_1;\theta)^\top\big]$$

est la matrice d'information de Fisher. Ainsi,

$$\hat{\theta}_n \overset{d}{\approx} \mathrm{N}\left(\theta,\, \frac{1}{n}\, I(\theta)^{-1}\right), \quad n \text{ grand.}$$

Intervalles de confiance

- Même si un estimateur est très précis, il donnera très rarement une estimation exactement juste (en fait, jamais dans le cas de modèles continus).
- Nous nous attendons néanmoins à ce qu'un bon estimateur “tombe dans le bon ordre de grandeur” avec une probabilité raisonnablement élevée.
- Dans la vie courante, on entend ou on lit souvent des phrases comme “le pourcentage de votes négatifs est estimé à 52% plus ou moins 3%”.

Les intervalles de confiance rendent ce type d'énoncés rigoureux.

Dans leur forme la plus simple (bilatérale et symétrique), ils visent à trouver une constante ε telle que le rang

$$\hat{\theta} \pm \varepsilon$$

ait une probabilité au moins égale à $1 - \alpha$ de contenir le vrai paramètre θ.

(α est typiquement petit, e.g. $\alpha = 0.05$, pour rendre $1 - \alpha$ grand)

Intervalle de confiance bilatéral

Soient $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, où $\theta \in \Theta \subseteq \mathbb{R}$, un échantillon aléatoire et $\alpha \in (0,1)$ une constante. Soient $L(X_1, \ldots, X_n)$ et $U(X_1, \ldots, X_n)$ deux statistiques, appelées respectivement la limite inférieure et la limite supérieure, telles que

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta \Big[L(X_1, \ldots, X_n) \leq \theta \leq U(X_1, \ldots, X_n) \Big] = 1 - \alpha.$$

Alors, l'intervalle aléatoire

$$\Big[L(X_1, \ldots, X_n)\,,\; U(X_1, \ldots, X_n) \Big],$$

est appelé un intervalle de confiance bilatéral pour θ avec un seuil de confiance $(1-\alpha)$.

Dans la diapositive précédente, nous avons considéré une forme particulière d'intervalle où L et U sont de la forme

$$L = \hat{\theta} - \varepsilon, \quad U = \hat{\theta} + \varepsilon.$$

Tous les intervalles ne sont pas nécessairement de cette forme, mais nous concentrerons notre attention sur ce type d'intervalles.

Intervalle de confiance unilatéral

Soient $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, où $\theta \in \Theta \subseteq \mathbb{R}$, un échantillon aléatoire et $\alpha \in (0,1)$ une constante. Soit $L(X_1, \ldots, X_n)$ une statistique telle que

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta \Big[L(X_1, \ldots, X_n) \leq \theta \Big] = 1 - \alpha.$$

Alors, l'intervalle aléatoire

$$\Big[L(X_1, \ldots, X_n)\,,\, +\infty \Big)$$

est appelé un intervalle de confiance unilatéral à gauche pour θ avec un seuil de confiance $(1-\alpha)$. De façon analogue, si $U(X_1, \ldots, X_n)$ satisfait

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta \Big[U(X_1, \ldots, X_n) \geq \theta \Big] = 1 - \alpha,$$

alors l'intervalle aléatoire

$$\Big(-\infty\,,\, U(X_1, \ldots, X_n) \Big]$$

est appelé un intervalle de confiance unilatéral à droite pour θ au seuil $(1-\alpha)$.

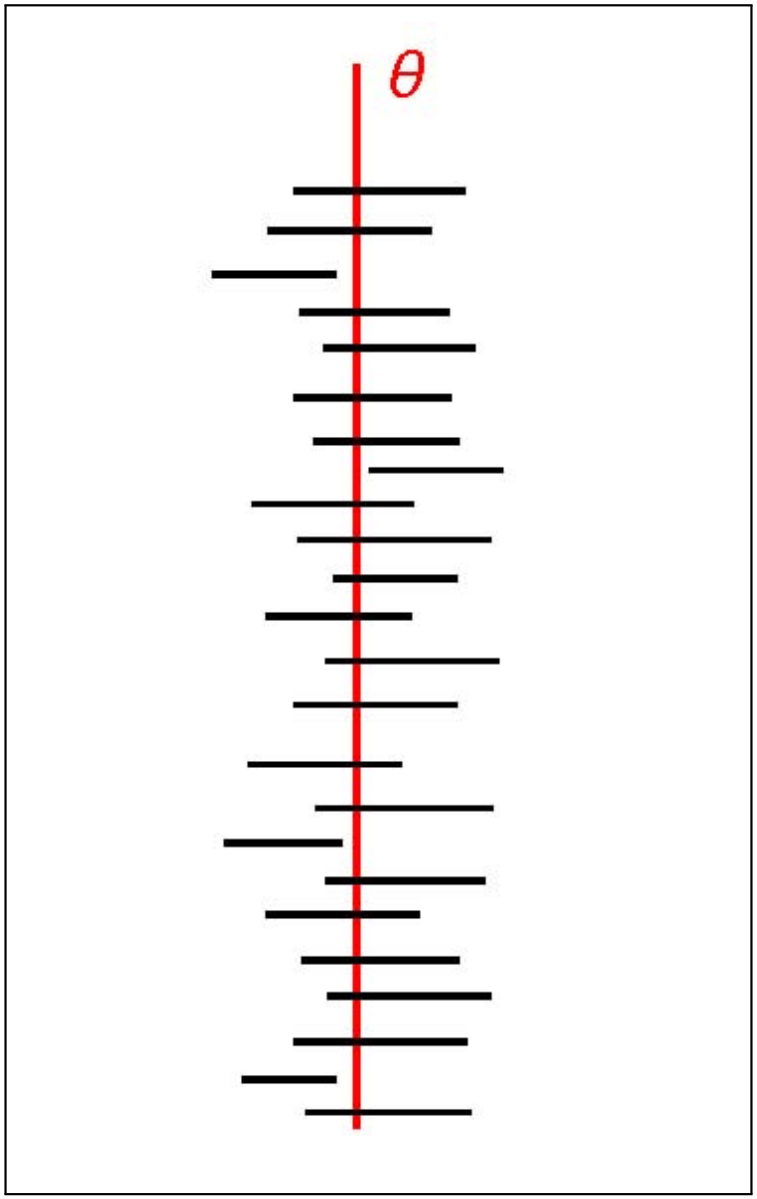
- Il faut faire attention lorsqu'on interprète un intervalle de confiance.
- Remarquez que

$$\inf_{\theta\in\Theta} \mathbb{P}_\theta\Big[L(X_1,\ldots,X_n) \leq \theta \leq U(X_1,\ldots,X_n)\Big] = 1-\alpha,$$

est une affirmation équivalente à

$$\inf_{\theta\in\Theta} \mathbb{P}_\theta\Big\{\theta \in \Big[L(X_1,\ldots,X_n)\,,\; U(X_1,\ldots,X_n)\Big]\Big\}\Big] = 1-\alpha.$$

- Toutefois, la deuxième façon d'écrire l'affirmation peut nous amener à une mauvaise interprétation de ce que signifie un intervalle de confiance.
- En effet, c'est l'intervalle $[L, U]$ qui est aléatoire et non le paramètre θ.
- Dire que "la probabilité que le paramètre tombe à l'intérieur de l'intervalle est au moins $1-\alpha$" est faux : le paramètre ne bouge pas, il est fixe!
- C'est l'intervalle qui peut changer pour différentes valeurs de l'échantillon $X_1, .., X_n$, et qui peut donc couvrir ou non le paramètre.
- Il faut donc dire "la probabilité que l'intervalle couvre le paramètre θ est au moins $(1-\alpha)$" .


θ

- Une façon différente de clarifier la situation est de remarquer que :

$$\mathbb{P}_\theta \Big[L(X_1, \ldots, X_n) \leq \theta \leq U(X_1, \ldots, X_n) \Big] =$$

$$= \mathbb{P}_\theta \Big[\{ L(X_1, \ldots, X_n) \leq \theta \} \} \cap \{ U(X_1, \ldots, X_n) \geq \theta \} \Big],$$

où le côté droit de l'expression accentue le fait que l'affirmation s'applique aux bornes aléatoires de confiance L et U, plutôt qu'au paramètre déterministe θ.

- Afin d'éviter toute confusion, il est préférable d'écrire $\mathbb{P}_\theta \{ [L, U] \ni \theta \}$ que $\mathbb{P}_\theta \{ \theta \in [L, U] \}$.

Example (Intervalle de confiance pour la moyenne d'une loi normale)

Soit $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, où μ est inconnu et σ^2 est connu. Nous voulons construire un intervalle bilatéral pour μ. Nous standardisons pour obtenir:

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

Ainsi, si $z_{\frac{\alpha}{2}}$ et $z_{1-\frac{\alpha}{2}}$ sont les $\alpha/2$ et $1-\alpha/2$ quantiles (respectivement) de la distribution $N(0,1)$, nous avons :

$$\mathbb{P}\left[z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\right] = 1 - \alpha.$$

En manipulant l'expression à l'intérieur de la probabilité, nous obtenons :

$$\mathbb{P}\left[z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

$$\begin{aligned}
&\iff \mathbb{P}\left[-\bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha \\
&\iff \mathbb{P}\left[\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha \\
&\iff \mathbb{P}\left[\bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha.
\end{aligned}$$

L'égalité ci-dessus est vraie quelque soit la vraie valeur de $\mu \in \mathbb{R}$. Donc si

$$L(X_1, \ldots, X_n) = \bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \qquad \& \qquad U(X_1, \ldots, X_n) = \bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},$$

alors $[L, U]$ est un intervalle de confiance au seuil $1-\alpha$. Par symétrie de $N(0,1)$,

$$\left[\ \underbrace{\bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}}_{L(X_1,\ldots,X_n)} \ , \ \underbrace{\bar{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}}_{U(X_1,\ldots,X_n)} \ \right]$$

Observez que l'intervalle est symétrique autour de $\bar{X}$, le EMV de μ. Pour mettre l'accent sur ce fait, on l'écrit souvent sous la forme

$$\boxed{\bar{X} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}}$$

Nous pouvons ainsi faire quelques observations importantes:

- La longueur de l'intervalle de confiance est $2z_{1-\alpha/2}\sigma/\sqrt{n}$, ce qui dépend de σ^2, n et α.
- Le paramètre σ^2 échappe à notre contrôle, puisque c'est la variance de la distribution $N(\mu, \sigma^2)$ sous-jacente.
- Nous pouvons cependant contrôler la taille de l'échantillon n et le seuil de confiance $1-\alpha$. En augmentant n, la longueur de l'intervalle est ré-échelonnée par un facteur de $1/\sqrt{n}$.
- D'un autre côté, diminuer α (i.e. augmenter la confiance $1-\alpha$) a pour effet d'augmenter la longueur de l'intervalle : plus nous voulons avoir de la confiance dans notre intervalle et plus l'intervalle sera grand (notons que la longueur de l'intervalle tend vers l'infini lorsque $\alpha \to 0$).

Maintenant, considérons le problème consistant à trouver un intervalle de confiance unilatéral à droite. En utilisant le fait que $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, nous pouvons écrire

$$\implies \mathbb{P}\left[\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha}\right] = 1-\alpha.$$

En manipulant l'expression, nous obtenons

$$\mathbb{P}\left[\bar{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \geq \mu\right] = 1-\alpha,$$

et l'intervalle

$$\left(\quad -\infty \quad , \quad \bar{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \quad \right].$$

est un intervalle de confiance unilatéral à droite avec au seuil $1-\alpha$. De façon similaire, un intervalle de confiance unilatéral à gauche avec un seuil $1-\alpha$ est donné par

$$\left[\quad \bar{X} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \quad , \quad +\infty \quad \right).$$

Example (Moyenne d'une distribution générale)

Soit $X_1, \ldots, X_n$ une collection de variables aléatoires iid de moyenne inconnue $\mu = \mathbb{E}[X]$ et de variance inconnue $\mathbb{E}[(X_1 - \mu)^2] = \sigma^2 < \infty$. On cherche un pivot approximatif afin de construire un intervalle pour μ.

- Par le théorème central limite, nous avons $\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2)$.
- Par la loi forte des grands nombres, $\sigma_n^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / n \xrightarrow{\mathbb{P}} \sigma^2$.

Maintenant, nous pouvons utiliser le théorème de Slustky afin de conclure que

$$\frac{\bar{X} - \mu}{\sigma_n / \sqrt{n}} \xrightarrow{d} Z \sim N(0,1),$$

On obtient, maintenant:

$$\mathbb{P}\left[\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma_n}{\sqrt{n}} \leq \mu \leq \bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma_n}{\sqrt{n}}\right] = \mathbb{P}\left[z_{\alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq z_{1-\alpha/2}\right]$$

$$\xrightarrow{n \to \infty} \mathbb{P}[z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}] = 1 - \alpha.$$

Qui donne l'intervalle approximatif $\bar{X} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma_n}{\sqrt{n}}$.

Qu'est-ce qui a permis aux exemples précédents de fonctionner ? Nous avons pu trouver un estimateur $\hat{\theta}$ de θ, ainsi qu'un estimateur de son écart-type $\hat{\sigma}_{\hat{\theta}}$, tels que

$$\frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}} \overset{d}{\approx} N(0, 1).$$

Cela conduit à l'intervalle de confiance (approximatif) de niveau $1 - \alpha$:

$$\hat{\theta} \pm z_{1-\frac{\alpha}{2}} \hat{\sigma}_{\hat{\theta}}.$$

Un tel intervalle est appelé intervalle de Wald, et il est valable chaque fois que nous disposons d'un TCL avec variance estimable pour notre estimateur $\hat{\theta}$.

Heureusement, sous des conditions de régularité sur le modèle de probabilité sous-jacent, nous connaissons un type d'estimateur qui satisfait précisément ces propriétés : l'estimateur du maximum de vraisemblance (EMV) :

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathrm{N}\left(0\,,\,\frac{1}{I(\theta)}\right) \implies \hat{\theta}_n \overset{d}{\approx} \mathrm{N}\left(\theta\,,\,\frac{1}{nI(\theta)}\right), \quad \text{pour } n \text{ grand.}$$

En résumé, nous obtenons un tableau général pour la construction d'intervalles de confiance (approximatifs), à condition que les hypothèses du TCL de l'EMV soient vérifiées :

Intervalles de confiance approximatifs de Wald via l'EMV

Confiance $\approx 1-\alpha$	$L(X_1,\ldots,X_n)$	$U(X_1,\ldots,X_n)$
Bilatéral	$\hat{\theta} - z_{1-\alpha/2}\frac{1}{\sqrt{I(\hat{\theta})n}}$	$\hat{\theta} + z_{1-\alpha/2}\frac{1}{\sqrt{I(\hat{\theta})n}}$
Unilatéral à gauche	$\hat{\theta} - z_{1-\alpha}\frac{1}{\sqrt{I(\hat{\theta})n}}$	$+\infty$
Unilatéral à droite	$-\infty$	$\hat{\theta} + z_{1-\alpha}\frac{1}{\sqrt{I(\hat{\theta})n}}$

Exercice : Utiliser la méthode delta pour construire une intervalle de confiance pour $\mathbb{P}[X_1 \leq x_0]$ lorsque $X_1, ..., X_n \overset{iid}{\sim} \text{Exp}(\lambda)$, pour $\lambda > 0$ inconnu et $x_0 > 0$ fixe.

Tests d'Hypothèse

Toute démarche scientifique procède selon le même schéma :

- énoncé d'une hypothèse capable d'être contredite par des données ;
- recolte des données (observées ou résultant d'une expérience);
- comparaison des données avec les prévisions de l'hypothèse;
- non-rejet, rejet ou modification eventuelle de l'hypothèse.

En des termes statistiques, dans le cadre d'un modèle, on itère les étapes suivantes:

- énoncé d'une hypothèse (sur les paramètres du modèle probabiliste)
  - cette hypothèse est capable d'être contredit par des données (utilisant une statistique, appellée statistique de test)
- recolte des données (observées ou résultant d'une expérience);
- rejet (ou non) l'hypothèse à partir de la comparaison entre les données et les résultats prédits par l'hypothèse.
  - Est-ce que cet écart est significatif?
    (c.-à-d. : résultat reproductible ou simple coïncidence fortuite ?)

Example (Recherche du boson de Higgs)

- Une des plus grandes questions du dernier quart de siècle en physique: savoir si le fameux *boson de Higgs* existait ou non.
- En utilisant le Modèle standard de la physique des particules, on peut calculer que le nombre moyen de diphotons produits s'il n'y avait pas de boson de Higgs serait au plus b.
- De façon similaire, si le boson de Higgs existait, ce nombre moyen dépasserait nettement b.
- Par des moyens de characterisation on sait que les événements correspondant à l'observation de diphotons suivent une distribution de Poisson avec une certain moyenne, disons μ.

Ainsi, l'hypothèse nulle (qui correspond à l'état de la nature si le boson de Higgs n'existait pas) est

$$H_0 : \mu \leq b,$$

et l'hypothèse alternative concurrente (qui décrit l'état de la nature si le boson de Higgs existait) est

$$H_1 : \mu > b$$

□

Example (Lancé d'une pièce de monnaie)

- Considérons une situation où nous voulons vérifier si une pièce de monnaie est équilibrée ou biaisée.
- Nous pouvons lancer la pièce n fois et enregistrer le résultat de chaque lancé.
- Nous souhaitons alors utiliser ces résultats afin de décider si la probabilité d'obtenir face est égale à $1/2$ ou différente de $1/2$.
- Nous ne somme pas vraiment interessés à savoir la valeur exacte: au lieu de concentrer nos efforts a déterminer la valeur précise, on veut utiliser l'échantillon de manière efficace pour décider si la pièce est équilibrée ou biaisée.
- Nous pourrions formaliser ce problème en disant que $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$ et que nous voulons décider entre l'hypothèse $H_0 : p \in \{\frac{1}{2}\}$ et l'hypothèse $H_1 : p \in (0,1) \setminus \{\frac{1}{2}\}$.

□

Afin de rendre les choses plus concrètes:

1. Soient deux ensembles Θ_0 et Θ_1, avec $\Theta_0 \cap \Theta_1 = \emptyset$. Il y a deux hypothèses scientifiques concurrentes pour un même phénomène :
   1. l'*hypothèse nulle* H_0 qui dit que $\theta \in \Theta_0$,
   $$H_0 : \theta \in \Theta_0,$$
   2. l'*hypothèse alternative* qui postule plutôt que $\theta \in \Theta_1$,
   $$H_1 : \theta \in \Theta_1.$$
2. Dans de nombreuses situations pratiques, les hypothèses prennent l'une des formes suivantes, pour un θ_0 donné :

$$\underbrace{\left\{ \begin{array}{l} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{array} \right\}}_{\text{“bilatéral”}} \quad \text{ou} \quad \underbrace{\left\{ \begin{array}{l} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{array} \right\} \text{ ou } \left\{ \begin{array}{l} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{array} \right\}}_{\text{“unilatéral”}}$$

3. Nous voulons utiliser l'échantillon $X_1, .., X_n$ que nous avons à disposition afin de décider à quel ensemble il appartient.

Opérationnellement, nous utiliserons l'échantillon pour choisir entre les deux hypothèses — un choix binaire.

On aura donc :

Fonction de test

Une fonction de test δ est une application mesurable $\delta : \mathcal{X}^n \to \{0, 1\}$.

On obtient 0 ou 1 selon que l'échantillon satisfait ou non une certaine condition :

$$\delta(X_1, \ldots, X_n) = \begin{cases} 1, & \text{si } T(X_1, \ldots, X_n) \in C, \\ 0, & \text{si } T(X_1, \ldots, X_n) \notin C, \end{cases}$$

où

- T est une statistique appelée statistique de test, et
- C est un sous-ensemble de l'image de T, appelé région critique.

De façon plus compacte :

$$\delta(X_1, \ldots, X_n) = 1\{T(X_1, \ldots, X_n) \in C\}$$

Dans la plupart des problèmes scientifiques, il existe une asymétrie naturelle :

- H_0 représente le statu quo, et H_1 la "nouvelle théorie".
- Par exemple, H_0 affirme qu'un "nouveau médicament n'a aucun effet", qu'"il n'existe pas de nouvelle particule fondamentale", ou "la pièce est équilibrée".

L'accent est donc mis sur le contrôle du risque de fausse découverte :

fausse découverte $\equiv$ erreur de type I $\equiv$ rejeter H_0 alors que H_0 est vraie.

Le risque de fausse découverte fait référence à la probabilité de rejeter à tort H_0.

Niveau de signification et tests respectant un niveau

Le niveau de signification $\alpha \in (0,1)$ est la probabilité maximale d'erreur de type I que l'on est prêt à tolérer. Une fonction de test δ respecte le niveau α si

$$\forall \theta_0 \in \Theta_0, \quad \mathbb{P}_{\theta_0}[\delta(X_1, \ldots, X_n) = 1] \leq \alpha.$$

Alors comment construire un test ? (= choisir statistique + région critique)

Considérons d'abord le cas bilatéral : $\left\{ \begin{array}{l} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{array} \right\}$

- Supposons disposer d'un IC $1-\alpha$, $I = [L, U]$.
- Un tel intervalle couvre le vrai paramètre avec probabilité $1-\alpha$.
- [juger avec confiance $1-\alpha$ si θ_0 est plausible] $\iff$ [vérifier si $\theta_0 \in I$]
- Plus précisément, si l'on rejette H_0 dès que $I \not\ni \theta_0$, i.e. $\delta = 1\{I \not\ni \theta_0\}$, alors :

$$\mathbb{P}_{\theta_0}[\delta = 1] = \mathbb{P}_{\theta_0}[I \not\ni \theta_0] = 1 - \mathbb{P}_{\theta_0}[I \ni \theta_0] = 1 - (1-\alpha) = \alpha.$$

Dans les cas unilatéraux, on applique la même logique : vérifier si la valeur frontière θ_0 appartient à l'intervalle unilatéral I.

(ou, de façon équivalente, si l'ensemble nul Θ_0 intersecte I ou non)

Rappelons qu'un estimateur qui satisfait $\dfrac{\hat{\theta}-\theta}{\hat{\sigma}_{\hat{\theta}}} \overset{d}{\approx} N(0,1)$ nous donne,

$\rightarrow$ un intervalle de confiance bilatéral $\approx 1-\alpha$ de bornes

$$\hat{\theta} \pm z_{1-\frac{\alpha}{2}} \hat{\sigma}_{\hat{\theta}},$$

$\rightarrow$ et un intervalle unilatéral $\approx 1-\alpha$ de borne (gauche/droite)

$$\hat{\theta} - z_{1-\alpha} \hat{\sigma}_{\hat{\theta}}, \quad \text{ou} \quad \hat{\theta} + z_{1-\alpha} \hat{\sigma}_{\hat{\theta}}.$$

Les tests bilatéraux/unilatéraux correspondents, appelés tests de Wald, sont :

- $\{H_0 : \theta = \theta_0\} : \delta = 1\{I \not\ni \theta_0\} = 1\left\{\frac{|\hat{\theta}-\theta_0|}{\hat{\sigma}_{\hat{\theta}}} > z_{1-\frac{\alpha}{2}}\right\}$
- $\{H_0 : \theta \leq \theta_0\} : \delta = 1\{(-\infty, \theta_0] \cap [L, +\infty) = \emptyset\} = 1\left\{\frac{\hat{\theta}-\theta_0}{\hat{\sigma}_{\hat{\theta}}} > z_{1-\alpha}\right\}$
- $\{H_0 : \theta \geq \theta_0\} : \delta = 1\{(-\infty, U] \cap [\theta_0, +\infty) = \emptyset\} = 1\left\{\frac{\theta_0-\hat{\theta}}{\hat{\sigma}_{\hat{\theta}}} > z_{1-\alpha}\right\}$

où nous avons effectivement renversé les étapes de la construction de l'IC.

En bref, la statistique de test est la distance standardisée (signée ou en valeur absolue) entre l'estimateur et "la frontière du nul", et la région critique est de la forme $(q, +\infty)$ pour un quantile approprié q.

Parfois la variance de l'estimateur est implicitement specifiée dans le cas bilatéral (par H_0), et donc il n'est pas strictement nécessaire de l'estimer :

Soient $X_1, \ldots, X_n \overset{iid}{\sim} \texttt{Exp}(\lambda)$. L'EMV et l'information de Fisher sont

$$\hat{\lambda} = 1/\bar{X} \qquad \& \qquad I(\lambda) = n/\lambda^2$$

et on peut vérifier la régularité. Alors dans ce cas

$$\hat{\sigma}_{\hat{\lambda}} = \frac{\hat{\lambda}}{\sqrt{n}}, \qquad \frac{\hat{\lambda}-\lambda}{\hat{\sigma}_{\hat{\lambda}}} \approx \mathrm{N}(0,1).$$

Ceci donne le test

$$\delta = 1\left\{ \frac{|\hat{\lambda}-\lambda_0|}{\hat{\lambda}/\sqrt{n}} > z_{1-\alpha/2} \right\}$$

pour $\{H_0 : \lambda = \lambda_0$ vs $H_1 : \lambda \neq \lambda_0\}$. Mais sous H_0, on peut aussi aussi prendre

$$\delta = 1\left\{ \frac{|\hat{\lambda}-\lambda_0|}{\lambda_0/\sqrt{n}} > z_{1-\alpha/2} \right\}.$$

en utilisant $\sigma^{(0)}_{\hat{\lambda}} = \lambda_0/\sqrt{n}$ sous H_0. Les deux versions sont presque equivalents pour n grand. Les deux formes peuvent être utilisées, mais la convention dans un test de Wald est d'utiliser $\hat{\sigma}_{\hat{\theta}}$.

Qu'en est-il de l'autre type d'erreur ?

- Erreur de type I $\equiv$ faux positif $\equiv$ fausse découverte
- Erreur de type II $\equiv$ faux négatif $\equiv$ manquer une véritable découverte

On préfère généralement parler de la puissance $\beta \in (0,1)$ d'un test plutôt que du risque de manquer une découverte :

$$\beta = 1 - \mathbb{P}\{\text{erreur de type II}\}.$$

On peut voir la puissance comme un indicateur de la sensibilité du test à détecter H_1 (analogie : un détecteur de fumée plus sensible capte plus vite une anomalie).

Qu'influence la puissance d'un test ? Regardons la perspective via les intervalles :

- On rejette lorsque $I \not\ni \theta_0$.
- Heuristiquement, plus l'intervalle est serré, plus il est probable que $I \not\ni \theta_0$
- (donc plus faibles sont les écarts à θ_0 que l'on peut espérer détecter.)

La puissance d'un test dépend directement de la longueur de l'intervalle de confiance associé : plus l'intervalle est serré, meilleure est la capacité du test à détecter des écarts à H_0.

On considère le test bilatéral au niveau α :

$$H_0 : \theta = \theta_0, \qquad \delta = 1\left\{ \frac{|\hat{\theta} - \theta_0|}{\hat{\sigma}_{\hat{\theta}}} > z_{1-\frac{\alpha}{2}} \right\}.$$

On suppose que la valeur vraie est

$$\theta_1 = \theta_0 - \varepsilon,$$

où $|\varepsilon|$ est appelé la "taille d'effet".

Alors, sous θ_1 :

$$\mathbb{P}_{\theta_1}\{\text{rejeter } H_0\} = \mathbb{P}_{\theta_1}\left\{ \frac{|\hat{\theta} - \theta_0|}{\hat{\sigma}_{\hat{\theta}}} > z_{1-\frac{\alpha}{2}} \right\} = \mathbb{P}_{\theta_1}\left\{ \frac{|(\hat{\theta} - \theta_1) - \varepsilon|}{\hat{\sigma}_{\hat{\theta}}} > z_{1-\frac{\alpha}{2}} \right\}.$$

Mais quand θ_1 est la vraie valeur du paramètre, notre estimateur vérifie

$$\underbrace{\frac{\hat{\theta} - \theta_1}{\hat{\sigma}_{\hat{\theta}}}}_{:=Z} \overset{d}{\approx} \mathrm{N}(0,1).$$

La puissance du test vaut donc approximativement

$$\begin{aligned}
\mathbb{P}_{\theta_1}\left\{\left|Z - \frac{\varepsilon}{\hat{\sigma}_{\hat{\theta}}}\right| > z_{1-\frac{\alpha}{2}}\right\} &= 1 - \mathbb{P}_{\theta_1}\left\{\frac{\varepsilon}{\hat{\sigma}_{\hat{\theta}}} - z_{1-\frac{\alpha}{2}} < Z < \frac{\varepsilon}{\hat{\sigma}_{\hat{\theta}}} + z_{1-\frac{\alpha}{2}}\right\} \\
&\approx 1 - \left[\Phi\left(\frac{\varepsilon}{\hat{\sigma}_{\hat{\theta}}} + z_{1-\frac{\alpha}{2}}\right) - \Phi\left(\frac{\varepsilon}{\hat{\sigma}_{\hat{\theta}}} - z_{1-\frac{\alpha}{2}}\right)\right] \\
&= \Phi\left(\frac{\varepsilon}{\hat{\sigma}_{\hat{\theta}}} - z_{1-\frac{\alpha}{2}}\right) + \left[1 - \Phi\left(\frac{\varepsilon}{\hat{\sigma}_{\hat{\theta}}} + z_{1-\frac{\alpha}{2}}\right)\right] \\
&= \Phi\left(\frac{\varepsilon}{\hat{\sigma}_{\hat{\theta}}} - z_{1-\frac{\alpha}{2}}\right) + \Phi\left(-\frac{\varepsilon}{\hat{\sigma}_{\hat{\theta}}} - z_{1-\frac{\alpha}{2}}\right).
\end{aligned}$$

L'intervalle de confiance bilatéral correspondant est :

$$I = \left[\hat{\theta} - z_{1-\frac{\alpha}{2}}\,\hat{\sigma}_{\hat{\theta}},\ \hat{\theta} + z_{1-\frac{\alpha}{2}}\,\hat{\sigma}_{\hat{\theta}}\right],$$

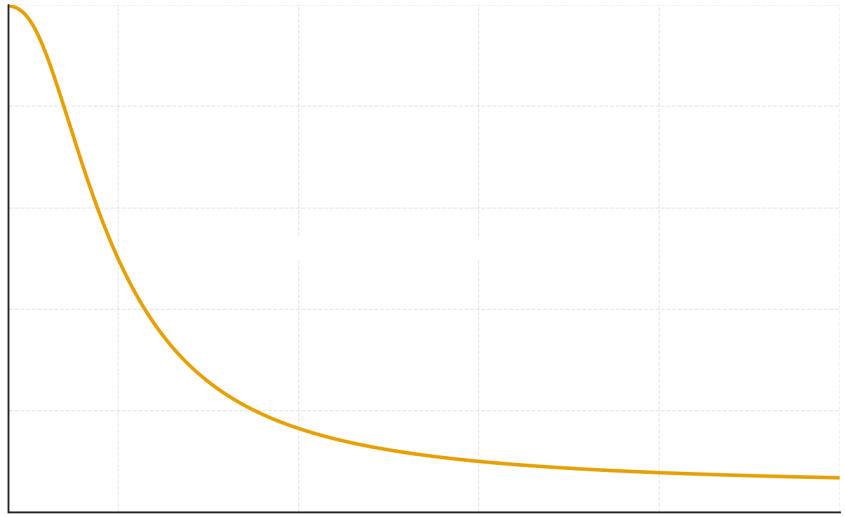
Sa longueur vaut :

$$|I| = 2z_{1-\frac{\alpha}{2}}\,\hat{\sigma}_{\hat{\theta}}, \quad \text{d'où} \quad \hat{\sigma}_{\hat{\theta}} = |I|/(2z_{1-\frac{\alpha}{2}}).$$

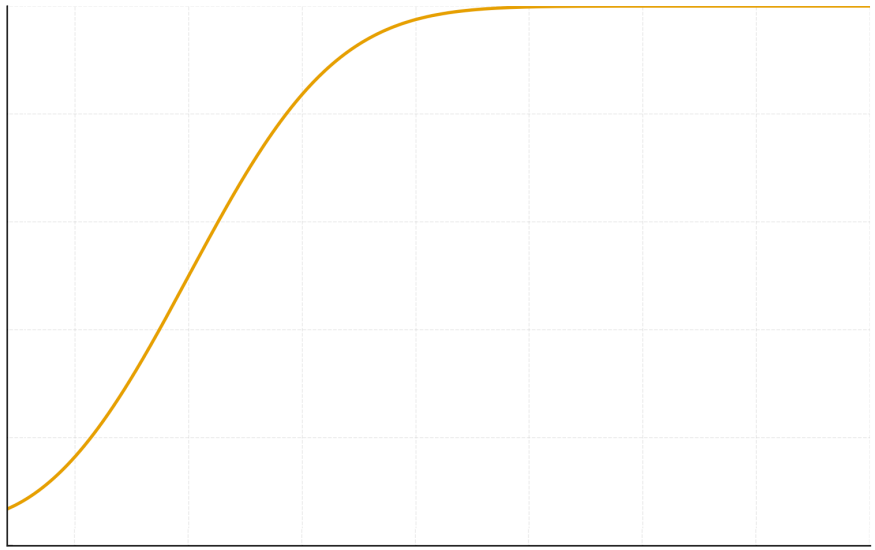
Alors la puissance peut être exprimée comme

$$\Phi\left(z_{1-\frac{\alpha}{2}}\left(\frac{2\varepsilon}{|I|} - 1\right)\right) + \Phi\left(-z_{1-\frac{\alpha}{2}}\left(\frac{2\varepsilon}{|I|} + 1\right)\right).$$

Illustration de la forme de la courbe de puissance en fonction de la longueur $|I|$



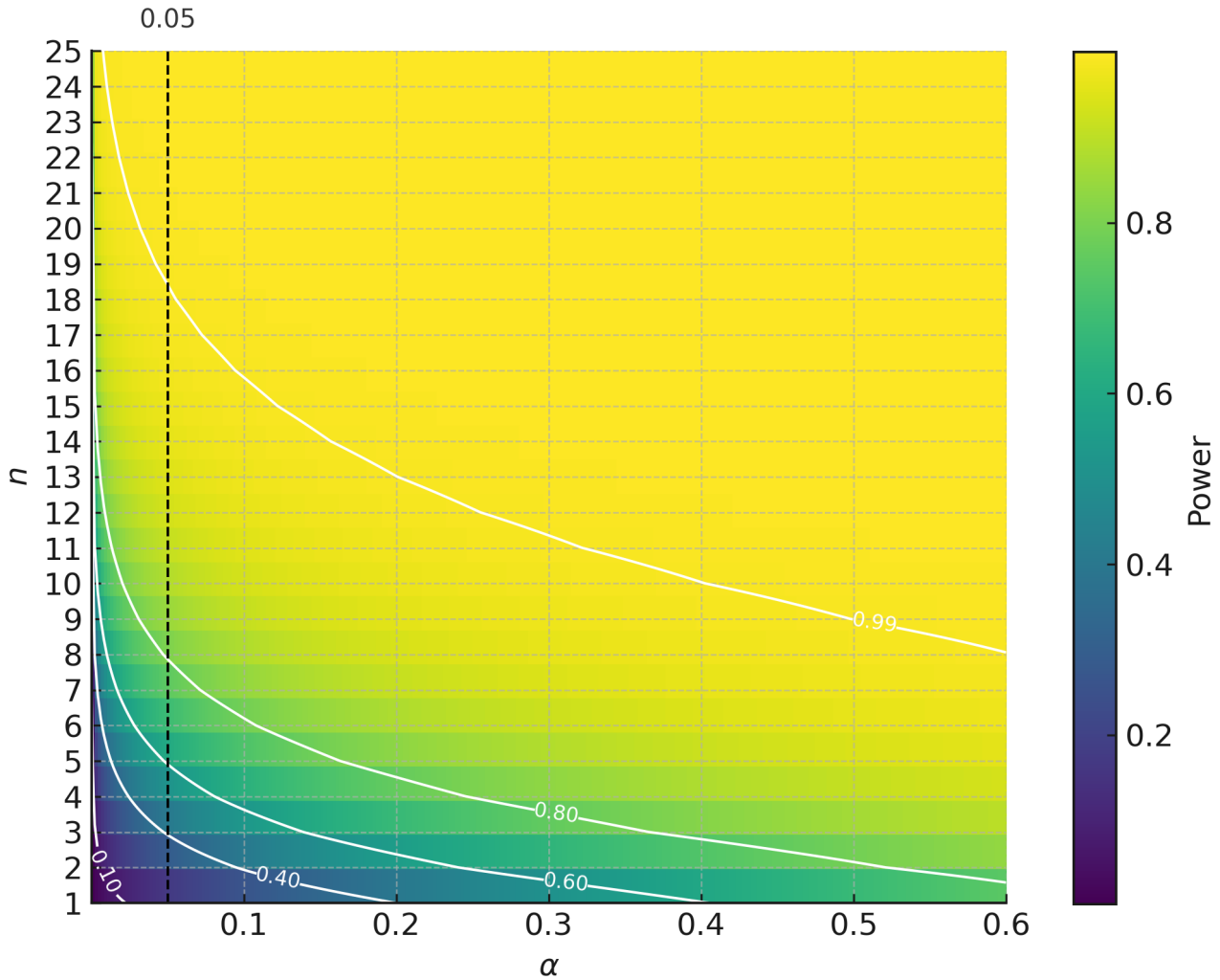
ou, de manière équivalente, en fonction de l'inverse de la longueur $1/|I|$



Rappelons que lorsque $\hat{\theta}$ est le maximum de vraisemblance, et sous des conditions de régularité,

$$\text{l'écart-type } \hat{\sigma}_{\hat{\theta}_n} \text{ (qui est } \propto |I|) \text{ décroît comme } \frac{1}{\sqrt{n}}.$$

Donc, pour une taille d'effet donnée ε (par exemple $\varepsilon = 1$), on peut tracer un diagramme de phase montrant comment la puissance dépend de n et de α :

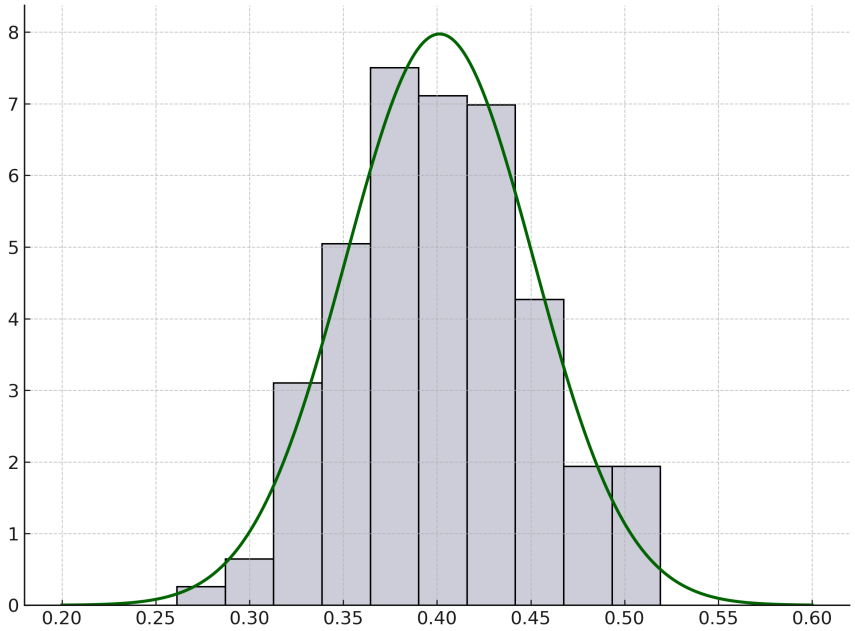


Un mot sur l'adéquation de l'ajustement

Peut-on tester l'hypothèse que notre choix de modèle est adéquat ?

- Supposons que nous utilisions un modèle F (avec densité f).
- Ayant des données $X_1, \ldots, X_n$ et construisons un histogramme à K classes.
- Comment tester si la "courbe s'ajuste à l'histogramme" ?

  (ou si la surface d'une densité conjointe s'ajuste à un histogramme planaire)



De tels tests sont appelés tests d'adéquation.

Dans l'une des premières contributions au domaine, Pearson a proposé d'utiliser

$$T_{\text{Pearson}} = \sum_{i=1}^{K} \frac{(N_i - np_i^{(0)})^2}{np_i^{(0)}}, \qquad \text{où} \quad p_i^{(0)} = \int_{C_i} f(u)\, du.$$

Cela ressemble à une "distance L^2 normalisée" entre histogramme et densité.

Sous l'hypothèse nulle $\{H_0 : X_i \overset{iid}{\sim} f\}$, on compare la valeur de la statistique au quantile $1-\alpha$ d'une loi χ^2_d, où :

1. $d = K - 1$ si la densité f est entièrement spécifiée (connue) ;
2. $d = K - 1 - q$ si nous avons dû estimer q paramètres de f.

Ce test est encore très utilisé aujourd'hui.

Expliquons (1) à l'aide de la théorie de la vraisemblance.
(la justification de (2) nécessite la théorie des rapports de vraisemblance.)

Exercice: Peut-on tester l'indépendance entre deux variables aléatoires de cette manière?

On observe une repartition dans les classes de l'histogramme :

$$N = (N_1, \ldots, N_K) \sim \text{Multinomial}\big(n; p_1, \ldots, p_K\big), \qquad p_i > 0, \sum_{i=1}^{K} p_i = 1.$$

Le paramètre a $K-1$ degrés de liberté. On utilise la parametrisation réduite

$$\theta = (\theta_1, \ldots, \theta_{K-1}) \in (0,1)^{K-1} \text{ avec } p_i(\theta) = \theta_i, \quad i < K, \text{ et } p_K(\theta) = 1 - \sum_{i=1}^{K-1} \theta_i.$$

On peut directement vérifier que l'EMV est donné par :

$$\hat{\theta}_i = \frac{N_i}{n}, \quad i = 1, ..., K-1.$$

On peut maintenent formuler notre objectif comme tester

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta^0. \end{aligned}$$

où $\theta_i^0 = \int_{C_i} f(u)\,du = p_i^{(0)}$, $i = 1, ..., K-1$.

Pour l'échantillon $(N_1, \ldots, N_K)$, la log-vraisemblance est

$$\ell(\theta) = \sum_{i=1}^{K-1} N_i \log \theta_i + N_K \log p_K(\theta).$$

La derivée partielle par composante i $(i = 1, \ldots, K-1)$ vaut :

$$\frac{\partial \ell(\theta)}{\partial \theta_i} = \frac{N_i}{\theta_i} - \frac{N_K}{p_K(\theta)}.$$

Alors :

$$-\mathbb{E}_\theta\left[\frac{\partial^2 \ell(\theta)}{\partial\theta\,\partial\theta^\top}\right] = n\left\{\operatorname{diag}\left(\frac{1}{\theta_1}, \ldots, \frac{1}{\theta_{K-1}}\right) + \frac{1}{p_K(\theta)}\,1 1^\top\right\} = nI(\theta),$$

où $1^\top = (1, ..., 1) \in \mathbb{R}^{K-1}$. On peut vérifier que $I(\theta)$ est inversible pour $\theta \in (0,1)^{K-1}$, d'inverse :

$$I(\theta)^{-1} = \left\{\operatorname{diag}(\theta_1, \ldots, \theta_{K-1}) - \theta\,\theta^\top\right\}.$$

Alors sous $H_0 : \theta = \theta^{(0)}$,

$$\sqrt{n}\big(\hat{\theta} - \theta^{(0)}\big) \xrightarrow{d} \mathrm{N}_{K-1}\big(0,\, I(\theta^{(0)})^{-1}\big) \overset{\textit{Slutsky}}{\Longrightarrow} \underbrace{\sqrt{n}\, I^{1/2}(\theta^{(0)})\big(\hat{\theta} - \theta^{(0)}\big)}_{:= T_n} \xrightarrow{d} \mathrm{N}_{K-1}\big(0,\, \mathrm{Id}\big).$$

On peut donc utiliser soit $|T_n|$ comme statistique du test, soit $W_n = T_n^2$:

$$W_n = n\,(\hat{\theta} - \theta^{(0)})^\top I_1(\theta^{(0)})(\hat{\theta} - \theta^{(0)}) \overset{d}{\approx} \chi^2_{K-1} \qquad \text{sous } H_0.$$

Si l'on pose $\Delta_i = \hat{\theta}_i - \theta_i^{(0)} = \frac{N_i}{n} - p_i^{(0)}$, $i < K$, on peut développer W_n :

$$W_n = n \sum_{i=1}^{K-1} \frac{\Delta_i^2}{p_i^{(0)}} \;+\; \frac{n}{p_K(\theta^{(0)})}\Big(\sum_{i=1}^{K-1} \Delta_i\Big)^2 \;=\; T_{\text{Pearson}}.$$

comme $p_K(\theta^{(0)}) = p_K^{(0)}$ et $\sum_{i=1}^{K-1} \Delta_i = p_K^{(0)} - N_K/n$,.

Postface sur l'Efficacité

Dans les trois problèmes statistiques que nous avons étudiés, la variance d'un estimateur (quasi) non biaisé a joué un rôle essentiel:

- elle détermine la précision de l'estimateur,
- elle fixe la longueur d'un intervalle de confiance,
- elle influence la puissance d'un test.

D'où des questions naturelles:

- Peut-on faire mieux, et utiliser les données de manière plus "efficace"? (obtenir davantage de précision avec la même quantité d'informations)?
- Existe-t-il une borne inférieure pour la variance d'un estimateur non biaisé?
- Peut-on l'atteindre?

Théorème (Cramér-Rao)

Soit $\hat{\theta}$ un estimateur non-biasé (c'est à dire $\mathbb{E}_\theta[\hat{\theta}] = \theta$) de variance finie. Supposons que le modèle est identifiable et la vraisemblance est suffisamment régulière pour permettre la double dérivation sous le signe intégral. Alors,

$$\mathrm{var}_\theta(\hat{\theta}) \geq \frac{1}{nI(\theta)},$$

où $I(\theta) = \mathbb{E}\left[-\partial_\theta^2 \log f(X_1;\theta)\right] = \mathbb{E}\left[\left(\partial_\theta f(X_1;\theta)\right)^2\right]$ est l'information de Fisher.

Rappelons que sous un peu plus de régularité, l'EMV vérifie

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathrm{N}\left(0\,,\,\frac{1}{I(\theta)}\right) \implies \hat{\theta}_n \overset{d}{\approx} \mathrm{N}\left(\theta\,,\,\frac{1}{nI(\theta)}\right), \quad \text{pour } n \text{ grand.}$$

Cela explique pourquoi nous avons mis autant l'accent sur l'inférence fondée sur la vraisemblance.

Preuve.

Rappelons que $\ell(\theta) = \sum_{i=1}^n \log f(X_i;\theta) = \sum_{i=1}^n \ell_i(\theta)$ et les identités de Bartlett,

$$\mathbb{E}_\theta[\ell'(\theta)] = 0, \qquad \mathrm{var}_\theta(\ell'(\theta)) = \sum_{i=1}^n \mathrm{var}_\theta(\ell_i'(\theta)) = nI(\theta).$$

Par dérivation sous le signe intégral et $\partial_\theta f = f\,\partial_\theta \log f$,

$$\mathrm{Cov}_\theta\big(\hat\theta, \ell'(\theta)\big) = \mathbb{E}_\theta[\hat\theta\ell'(\theta)] - \mathbb{E}_\theta[\hat\theta]\underbrace{\mathbb{E}_\theta[\ell'(\theta)]}_{=0} = \mathbb{E}_\theta\big[\hat\theta\,\ell'(\theta)\big] = \partial_\theta \mathbb{E}_\theta[\hat\theta] = \partial_\theta\theta = 1.$$

Par l'inégalité de corrélation (le fait que $|\mathrm{corr}| \le 1$),

$$\underbrace{\mathrm{Cov}_\theta(\hat\theta, \ell'(\theta))^2}_{=1} \le \mathrm{Var}_\theta(\hat\theta)\,\underbrace{\mathrm{Var}_\theta(\ell'(\theta))}_{nI(\theta)}.$$

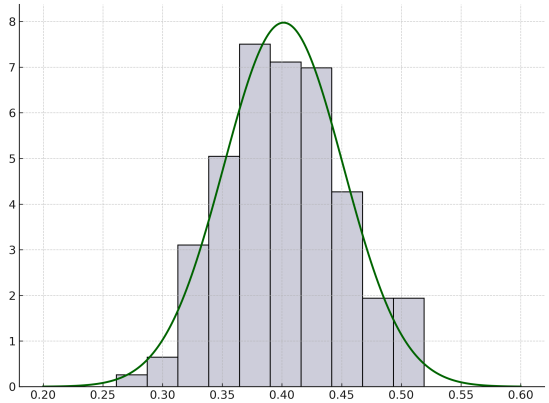
□

Matériel bonus (non examinable)

On considère un histogramme à classes régulières $(C_i)_{i\in\mathbb{Z}}$ de largeur $h > 0$.

À partir de données $X_1, \dots, X_n \overset{iid}{\sim} f$, on définit

$$\hat{f}_h(x) = \sum_i \hat{f}_i \, 1_{C_i}(x), \qquad \hat{f}_i = \frac{N_i}{nh}, \quad N_i = \sum_{k=1}^{n} 1\{X_k \in C_i\}.$$



On s'intéresse à l'erreur quadratique intégré (IMSE) :

$$\mathbb{E}\left[\|\hat{f}_h - f\|_2^2 \right] = \mathbb{E}\left[\int (\hat{f}_h(x) - f(x))^2 \, dx \right].$$

Objectif : comprendre le rôle de h et le taux de convergence.

Comme $\hat{f}_h$ est constante sur chaque classe C_i,

$$\int (\hat{f}_h - f)^2 = \sum_i \int_{C_i} (\hat{f}_i - f(x))^2 \, dx.$$

On introduit la "valeur moyenne de f" sur la classe :

$$f_i = \frac{1}{h} \int_{C_i} f(u) \, du, \qquad \mathbb{E}[\hat{f}_i] = f_i.$$

En écrivant $\hat{f}_i - f = (\hat{f}_i - f_i) + (f_i - f)$, on obtient, pour chaque i,

$$\int_{C_i} (\hat{f}_i - f)^2 = h(\hat{f}_i - f_i)^2 + \int_{C_i} (f_i - f)^2,$$

le terme croisé s'annulant exactement. Prenant l'espérance et en sommant sur i :

$$\mathbb{E}\|\hat{f}_h - f\|_2^2 = \underbrace{\sum_i h \operatorname{Var}(\hat{f}_i)}_{\text{fluctuations statistiques}} + \underbrace{\sum_i \int_{C_i} (f_i - f)^2}_{\text{biais d'approximation}}.$$

On traite séparément les deux termes.

Pour chaque classe C_i,

$$N_i \sim \text{Binomial}(n, p_i), \qquad p_i = \int_{C_i} f(u)\, du.$$

On a alors

$$\text{Var}(\hat{f}_i) = \frac{p_i(1-p_i)}{nh^2}.$$

Donc

$$\sum_i h\, \text{Var}(\hat{f}_i) = \frac{1}{nh} \sum_i p_i(1-p_i) \;\leq\; \frac{1}{nh} \sum_i p_i = \frac{1}{nh}.$$

Le terme de variance est d'ordre $(nh)^{-1}$.

Supposons que la densité f soit Lipschitz (par exemple C^1 de dérivée bornée):

$$|f(x) - f(y)| \leq L|x - y|.$$

Alors, pour $x \in C_i$,

$$|f(x) - f_i| \leq \frac{1}{h} \int_{C_i} |f(x) - f(u)| \, du \leq Lh.$$

Il en résulte

$$\int_{C_i} (f_i - f)^2 \;\leq\; C\, h^2 \int_{C_i} f(x)\, dx = C\, h^2\, p_i,$$

pour une constante universelle C.

En sommant sur i :

$$\sum_i \int_{C_i} (f_i - f)^2 \;\leq\; C\, h^2.$$

Le biais intégré est d'ordre h^2.

On a donc la borne globale

$$\mathbb{E}\|\hat{f}_h - f\|_2^2 \;\leq\; \frac{1}{nh} \;+\; Ch^2.$$

Sous les conditions asymptotiques usuelles :

$$h \to 0, \qquad nh \to \infty,$$

les deux termes tendent vers 0. Interprétation :

- $[h \to 0]$ On estime une fonction non constante par une moyenne locale ; il faut donc moyenner sur des régions de plus en plus petites.
- $[nh \to \infty]$ Pour que le procédé de moyennage fonctionne, il faut disposer d'un nombre de plus en plus grand d'observations par classe

En équilibrant biais et variance :

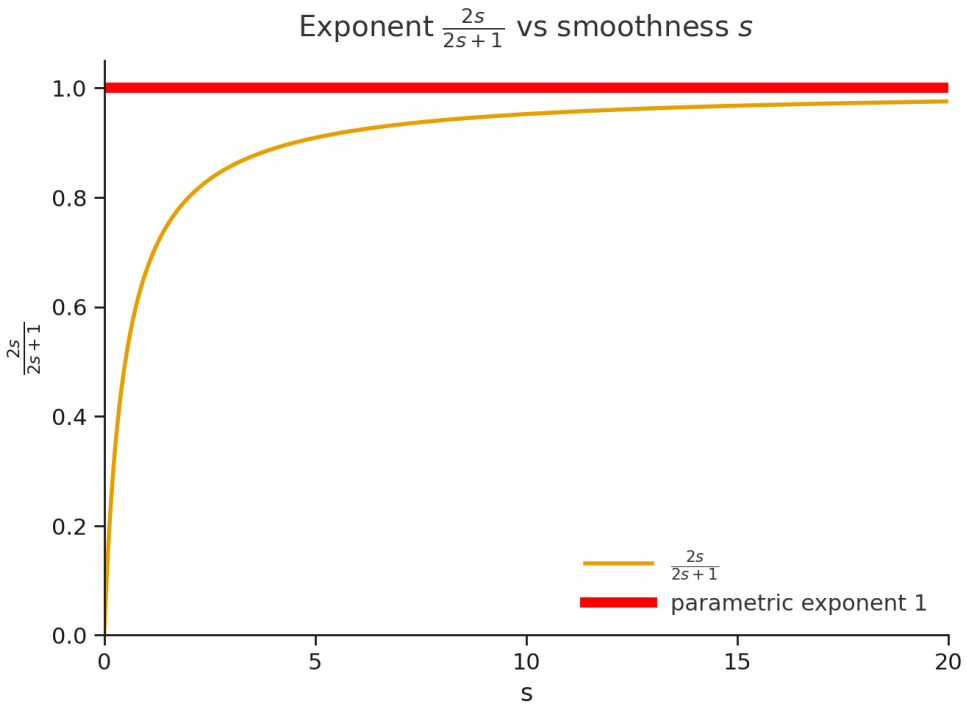
$$h^2 \asymp \frac{1}{nh} \quad \Longrightarrow \quad h \asymp n^{-1/3}.$$

On obtient alors le taux

$$\boxed{\mathbb{E}\|\hat{f}_h - f\|_2^2 = O(n^{-2/3})}.$$

- Le taux $n^{-2/3}$ reflète un compromis :
  - classes fines $\Rightarrow$ bon biais mais peu d'observations ;
  - classes larges $\Rightarrow$ faible variance mais mauvaise localisation.

- Ce cadre sert de prototype pour des estimateurs plus lisses (noyaux, projections,...) qui peuvent atteindre un taux de $n^{-4/5}$.
- Si f est C^s de s-dérivée bornée, on peut améliorer le taux à $n^{-\frac{2s}{2s+1}}$



On observe $X_1, \dots, X_n \overset{iid}{\sim} f$ sur $\mathbb{R}^d$.

On partitionne $\mathbb{R}^d$ en classes hypercubiques $(C_i)_{i \in \mathbb{Z}^d}$ de côté h :

$$\text{Vol}(C_i) = h^d.$$

On définit l'histogramme d-dimensionnel

$$\hat{f}_h(x) = \sum_{i \in \mathbb{Z}^d} \hat{f}_i \, 1_{C_i}(x), \qquad \hat{f}_i = \frac{N_i}{nh^d}, \quad N_i = \sum_{k=1}^{n} 1\{X_k \in C_i\}.$$

Notons

$$p_i = \int_{C_i} f(u)\, du, \qquad f_i = \frac{1}{h^d} \int_{C_i} f(u)\, du = \mathbb{E}[\hat{f}_i].$$

Même idée : moyenne locale, mais la "taille" d'une classe est désormais h^d.

Comme en dimension 1,

$$\mathbb{E}\|\hat{f}_h - f\|_2^2 = \sum_i h^d \operatorname{Var}(\hat{f}_i) \; + \; \sum_i \int_{C_i} (f_i - f(x))^2 \, dx.$$

Variance. $N_i \sim \text{Binomial}(n, p_i)$ donc

$$\operatorname{Var}(\hat{f}_i) = \frac{p_i(1-p_i)}{nh^{2d}} \quad \Longrightarrow \quad \sum_i h^d \operatorname{Var}(\hat{f}_i) \leq \frac{1}{nh^d}.$$

Biais (approximation). Si f est Lipschitz sur $\mathbb{R}^d$ (pour la norme euclidienne), alors sur un cube de côté h on a typiquement $|f(x) - f_i| \lesssim h$, donc

$$\sum_i \int_{C_i} (f_i - f)^2 = O(h^2).$$

Ainsi,

$$\boxed{\mathbb{E}\|\hat{f}_h - f\|_2^2 \; \lesssim \; h^2 + \frac{1}{nh^d}.}$$

En équilibrant

$$h^2 \asymp \frac{1}{nh^d} \quad \Longrightarrow \quad h \asymp n^{-1/(d+2)}.$$

On obtient le taux IMSE

$$\boxed{\mathbb{E}\|\hat{f}_h - f\|_2^2 = O\left(n^{-2/(d+2)}\right).}$$

Interprétation (malédiction de la dimension).

- Le nombre de classes nécessaires pour une résolution h est de l'ordre de h^{-d} (croît exponentiellement en d).
- La condition "assez d'observations par classe" devient

  $$nh^d \to \infty,$$

  beaucoup plus exigeante lorsque d grand.
- Pour viser une précision ε (IMSE $\approx \varepsilon$), le taux donne typiquement

  $$n \approx \varepsilon^{-(d+2)/2},$$

  donc un coût en taille d'échantillon qui explose avec d.

Un langage cohérent pour l'incertitude: modéliser, quantifier, et inférer.

- Fondations. À partir des ensembles et des σ-algèbres, les axiomes de Kolmogorov définissent un cadre mathématique rigoureux pour la probabilité, les variables aléatoires et leurs lois (fonctions de répartition).
- Modèles. Les lois de probabilité sont des *hypothèses structurées* sur les données. Les moments, l'entropie et les familles exponentielles traduisent des contraintes physiques en modèles probabilistes.
- Limites et fluctuations. La loi des grands nombres et le théorème central limite montrent que de propriétés "lisses et symétriques" convergent avec des fluctuations d'ordre $\sqrt{n}$ universelles gaussiennes à grande échelle.
- Inférence statistique. À partir des données, on estime (EMV), on quantifie l'incertitude (intervalles de confiance) et on teste des hypothèses (tests de vraisemblance, χ^2), avec des limites fondamentales (Cramér–Rao).
- Choix du modèle. Lorsque l'on dispose d'une *théorie physique solide*, un modèle paramétrique est naturel : peu de paramètres, interprétables, et une inférence statistique efficace. À l'inverse, lorsque la structure du phénomène est incertaine, les méthodes non paramétriques permettent d'apprendre *une partie de la forme du modèle elle-même* à partir des données, au prix d'un compromis biais–variance

Ttoute inférence implique un arbitrage raisonné entre théorie, modélisation, et données disponibles.