

Je ne comprenais rien aux probabilités et aux statistiques mais
maintenant ça va mieux !

François Schwarzentruher
ENS Rennes

Version du 23 mars 2024 - version provisoire

Table des matières

1	Avant-propos	9
1.1	Remerciement	9
1.2	Bibliographie	9
1.3	Motivation	9
I	Probabilités	11
2	Théorie des probabilités	13
2.1	Ensemble des mondes possibles	13
2.2	Événements	13
2.3	Tribu	13
2.4	Tribu borélienne	14
2.5	Mesure de probabilité	14
2.6	Espace probabilisé	15
2.7	Indépendance	16
3	Probabilités conditionnelles	17
3.1	Définition	17
3.2	Formule de Bayes	18
3.3	Indépendance conditionnelle	20
4	Variables aléatoires et lois	23
4.1	Variable aléatoire	23
4.2	Indépendance	23
4.2.1	Deux variables	23
4.2.2	Plusieurs variables	24
4.3	Lois discrètes	24
4.3.1	Dirac	24
4.3.2	Loi de Bernoulli	25
4.3.3	Loi binomiale	25
4.3.4	Loi géométrique	25
4.3.5	Lois de Poisson	26
4.4	Lois continues (à densité)	27
4.4.1	Lois uniformes	27
4.4.2	Lois exponentielles	27
4.4.3	Lois normales	28
4.5	Densité marginale	29
4.6	Fonction de répartition	30
4.7	Lois mélanges (mixture)	31
5	Propriétés sur les distributions	33
5.1	Espérance	33
5.1.1	Théorème de transfert	33
5.1.2	Formules dans le cas discret et continue	34
5.1.3	Propriétés	35
5.1.4	Technique de la fonction test	35

5.2	Variance	36
5.2.1	Propriétés	36
5.3	Covariance	37
5.4	Variance d'une somme	38
5.5	Inégalités	39
5.5.1	Inégalité de Markov	39
5.5.2	Inégalité de Bienaymé-Tchebychev	39
5.5.3	Inégalité de Jensen	39
6	Simulation de lois	41
6.1	Loi uniforme sur $[0, 1]$	41
6.2	Méthode d'inversion	41
6.2.1	Cas où F est inversible	41
6.2.2	Cas général	42
6.3	Loi normale	43
6.4	Méthode du rejet	44
6.4.1	Cas simple : loi uniforme sur le disque unité	44
6.4.2	Loi uniforme sur le cercle unité	46
6.4.3	Algorithme de rejet pour une fonction de densité à support compact et majorée	46
6.4.4	Algorithme du rejet généralisé	47
7	Convergence de suites de v.a.	51
7.1	Notation pour la moyenne	51
7.2	Différents types de convergence	51
7.2.1	Converge presque sûre	51
7.2.2	Convergence en probabilité	51
7.2.3	Convergence en loi	52
7.2.4	Comparaison des types de convergence	53
7.3	Théorème de Slutsky	53
7.4	Loi des grands nombres	54
7.5	Théorème central limite	55
7.6	D'autres convergences en loi	56
7.6.1	Loi des événements rares	56
7.6.2	Comparaison loi géométrique et loi exponentielle	56
8	Vecteurs aléatoires	57
8.1	Espérance	57
8.2	Matrice de variance-covariances	57
8.2.1	Définition	57
8.2.2	Matrice symétrique	58
8.2.3	Matrice définie semi-positive	58
8.3	Loi multinomiale	59
8.4	Lois normales vectorielles	62
8.4.1	Définition	62
8.4.2	Caractérisation par espérance et matrice de covariance	63
8.4.3	Densité quand la matrice de covariance est définie strictement positive	63
8.4.4	Changement affine	66
8.5	Théorème centrale limite vectoriel	66
II	Statistiques	69
9	Introduction	71
9.1	Modèles statistiques	71
9.2	Échantillon aléatoire	71

10 Estimation	73
10.1 Exemples	73
10.2 Estimateur	73
10.3 Biais	74
10.3.1 Estimateur sans biais	74
10.3.2 Estimateur asymptotiquement sans biais	74
10.3.3 Estimateurs de la variance	74
10.4 Maximum de vraisemblance	76
10.4.1 Vraisemblance	76
10.4.2 Maximiser la vraisemblance	76
10.4.3 Exemples	77
10.5 Méthode des moments (*)	78
11 Clustering	79
11.1 Applications	79
11.2 Clusters	79
11.3 Algorithme des k -means	80
11.3.1 Problème	80
11.3.2 Algorithme de Lloyd	80
11.3.3 Optimisations	80
11.4 Clustering par modèle de mélange gaussien	80
11.4.1 Maximum de vraisemblance	81
11.4.2 Difficulté de calcul	81
11.4.3 Variables latentes	81
11.4.4 Calcul de l'espérance par rapport aux variables latentes	81
11.4.5 Calcul de $\mathbb{P}(Z = k \mid X = x_i)$	82
11.4.6 Estimation des paramètres maximisant cette espérance	82
11.4.7 Pseudo-code	82
11.5 Relation avec K-moyennes	83
12 Lois du χ^2	85
12.1 Définition	85
12.2 Propriétés	85
12.3 Théorème de Cochran	85
13 Intervalles de confiance	87
13.1 Mon premier exemple	87
13.2 Définition	87
13.3 Quantiles	88
13.4 Estimation de la moyenne d'une loi normale avec écart-type connu	89
13.4.1 Statistique pivotale	89
13.4.2 Intervalle de confiance	90
13.5 Estimation de la moyenne d'une loi normale avec écart-type inconnu	91
13.5.1 Se débarrasser de σ	91
13.5.2 Statistique pivotale	91
13.5.3 Intervalle de confiance	93
13.6 Intervalle de confiance asymptotique	93
13.6.1 Moyenne inconnu, écart-type connu	93
13.6.2 Sondage	93
13.7 Intervalle de confiance pour un quantile	93
13.8 Intervalle de prédiction	95
13.8.1 Cas général iid	95
13.8.2 Cas normal iid	96

14 Généralités sur les tests statistiques	99
14.1 Hypothèses	99
14.2 Algorithme	100
14.3 Exemple : pièce équilibrée ou pipée ?	100
14.4 Intervalle de fluctuation	101
14.5 Test z	101
14.6 Test de Student	102
14.7 Métaphore du procès	103
14.7.1 Le procès	103
14.7.2 Erreurs de jugement	104
14.7.3 Compromis entre erreurs de type 1 et 2	104
14.7.4 Démarche commune	104
14.8 Test t de Student pour échantillons indépendants	105
14.9 Test t de Welsh pour échantillons indépendants	105
14.10 Test de Kolmogorov-Smirnov	105
14.11 Méthode de Neyman et Pearson (*)	106
15 Tests du χ^2	107
15.1 Motivation	107
15.2 Test d'adéquation à une loi donnée	107
15.2.1 Statistique utilisée	108
15.2.2 Comportement de la statistique quand H_0 fausse	109
15.2.3 Comportement de la statistique quand H_0 vraie	109
15.2.4 Algorithme	110
15.3 Test d'adéquation à une famille de lois	111
15.3.1 Statistique	112
15.3.2 Algorithme	112
15.4 Test d'indépendance	113
15.4.1 Statistique utilisée	113
15.4.2 Algorithme	113
16 Régression linéaire	115
16.1 Idée informelle	115
16.2 Exemples d'applications	116
16.3 Définition de la régression linéaire	116
16.4 Régression linéaire simple	117
16.5 Formulation matricielle	118
16.6 Solution	118
16.7 Théorème de Gauss-Markov	119
16.8 Lien avec le maximum de vraisemblance	120
16.9 Qualité d'une régression	120
16.9.1 Tester qu'un coefficient est nul	121
17 Régression logistique	123
17.1 Idée informelle	123
17.2 Exemples d'application	124
17.3 Définition	124
18 Classification bayésienne naïve	125
18.1 Bayes est gentil	125
18.2 Naïveté	126
18.3 Estimations	126
18.4 Limite du modèle	126

III	Partie théorique	127
19	Fonctions caractéristiques	129
19.1	Définition	129
19.2	Propriétés	129
19.3	Caractérisation	130
19.4	Famille tendue de mesures	130
19.5	Théorème de Levy	131
19.6	Démonstration du théorème centrale limite	132

Chapitre 1

Avant-propos

1.1 Remerciement

Je remercie Louis Thiry, Emmanuelle Becker, Romaric Gaudel, Benoît Cadre, Elisa Fromont, Constance Bocquillon, Lendy Mulot pour les discussions.

Je remercie Nemokary qui a réalisé certaines illustrations de ce polycopié.

1.2 Bibliographie

Ces notes de cours proviennent de différentes lectures. La première partie provient de [GK19], [BL21]. La deuxième partie sur les statistiques est un mélange de [Aze22], [HTFF09], [Lec02].

1.3 Motivation

Plus tard, vous serez **informatien** ou **informaticienne**. Quelque soit votre domaine de recherche de prédilection, vous aurez besoin de probabilités et/ou de statistiques avec une forte probabilité!

Algorithmique Algorithmes probabilistes. Structures de données probabilistes. Méthode probabiliste.

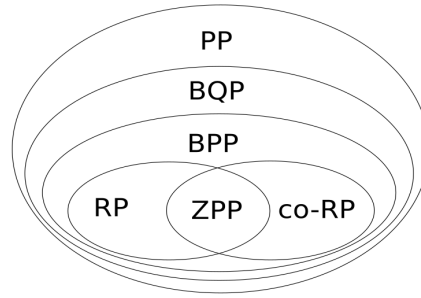
Apprentissage. La plupart des modèles d'apprentissage sont des modèles statistiques. On cherche à trouver des paramètres pour coller aux données (maximum de vraisemblance).



Bioinformatique. Pareil. Tests statistiques.

Cryptographie. Propriété de sécurité probabiliste. Assistant de preuve EasyCrypt <https://github.com/EasyCrypt/easycrypt>

Informatique théorique Théorie de la complexité avec des classes probabilistes. Théorème PCP (Probabilistically checkable proof).



Interface humain-machine. Contacter Lendy Mulot (équipe RAINBOW). On donne des tâches comme ramasser un objet dans un environnement virtuel. Question : est-ce que le feedback haptique (i.e. feedback via le toucher) est utile ? Autre question : quelle est l'intensité idéale pour vibrer une manette de jeux vidéos ?

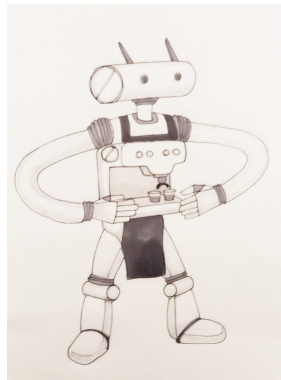


Langage de programmation. Modèle de calcul probabiliste (Christine Tasson). Langage de programmation synchrone probabiliste (par exemple Probelus).

Logique mathématique. On peut être amené à faire du raisonnement probabiliste. Citons probabilistic logic, la logique floue (qui repose sur la théorie des possibilités, qui étend la théorie des probabilités)

Représentation des connaissances. Citons les réseaux bayésiens.

Robotique. Il y a des probabilités dans les MDP, POMDP, decPOMDP.



Psychologie. Est-ce qu'une méthode (formation, etc.) augmente l'estime de soi ?

Vérification de programmes Citons les logiques PLTL (probabilistic linear temporal logic), PCTL (probabilistic computation tree logic). Citons aussi le model checking statistique.

Première partie

Probabilités

Chapitre 2

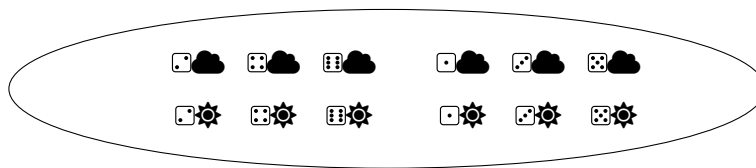
Théorie des probabilités

Les probabilités sont utilisées pour mesurer de combien un certain événement, par exemple, un dé donne 4, est probable. Typiquement, pour un dé non pipé, la probabilité que le dé donne 4 est de $\frac{1}{6}$. Les bayésiens donnent aussi des probabilités à des théories. Par exemple, la probabilité que le dé soit non pipé est de $\frac{2}{5}$. Mais avant de parler d'événements, on considère des mondes possibles.

2.1 Ensemble des mondes possibles

On note Ω un ensemble non vide de mondes possibles. Un monde possible dit exactement tout sur tout : combien un dé va donner, quelle température il fait à Rennes demain, si une théorie est vraie ou non (pour les bayésiens), etc.

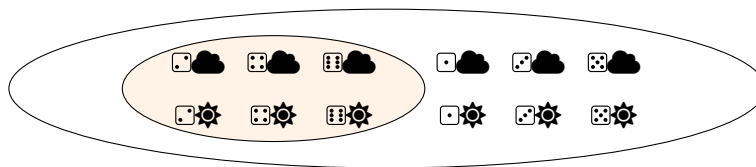
Exemple 1. Voici un exemple d'ensemble Ω avec 12 mondes possibles :



2.2 Événements

Définition 2. Un **événement** (ou événement aléatoire) est un sous-ensemble de Ω .

Exemple 3. $A :=$ 'Le dé affiche un nombre pair'.



! Un événement est un sous-ensemble de Ω et non un point ω de Ω . Enfin, un singleton $\{\omega\}$, s'il est mesurable est un événement.

2.3 Tribu

Comme l'on va attribuer une probabilité qui mesure de combien un événement A est probable, on fait appel à la **théorie de la mesure**. Tous les sous-ensembles ne sont pas forcément mesurables. Dans la théorie de la mesure, les sous-ensemble mesurable font parti de ce que l'on appelle une **tribu**. Les événements 'licites' sont donc les sous-ensembles d'une tribu.

Définition 4. Une **tribu** \mathcal{F} est un ensemble de sous-ensembles de Ω avec :

1. $\emptyset, \Omega \in \mathcal{F}$;
2. \mathcal{F} est stable par complémentaire, union et intersection dénombrables.

2.4 Tribu borélienne

La tribu borélienne, que l'on note ici $Borel(\mathbb{R})$, nommée en hommage à Émile Borel, est celle qui est généralement utilisée quand $\Omega = \mathbb{R}$. Voici comment elle est construite. On suppose que les intervalles ouverts $] - \infty, a[$ sont dans $Borel(\mathbb{R})$. Puis, on suppose $Borel(\mathbb{R})$ stable par complémentaire, union, et intersection dénombrables. Du coup :

- $[a, +\infty[= \mathbb{R} \setminus] - \infty, a[$ est dans $Borel(\mathbb{R})$;
- $]a, +\infty[= \bigcup_{n \in \mathbb{N}} [a + 1/n, +\infty[$ est dans $Borel(\mathbb{R})$;
- $[a, b] = [a, +\infty[\cap] - \infty, b]$ est dans $Borel(\mathbb{R})$;
- $\{a\} = [a, a]$ est dans $Borel(\mathbb{R})$;
- Tout ensemble dénombrable $X = \bigcup_{x \in X} \{x\}$ est dans $Borel(\mathbb{R})$.

Les deux définitions suivantes formalisent exactement cette construction.

Définition 5 (tribu engendrée). *Soit C un ensemble de sous-ensembles de Ω . La tribu engendrée par C est la plus petite tribu contenant C .*

Définition 6. La **tribu borélienne** $Borel(\mathbb{R})$ est la tribu engendrée par l'ensemble des intervalles ouverts de \mathbb{R} .

2.5 Mesure de probabilité

Nous sommes maintenant prêt à définir une mesure de probabilité : c'est une fonction \mathbb{P} , qui à tout événement A , associe un nombre $\mathbb{P}(A)$ entre 0 et 1, qui mesure de combien l'événement est probable. Le nombre $\mathbb{P}(A)$ est la probabilité que A soit vraie.

Définition 7. Une **mesure de probabilité** est une application de $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ vérifiant :

- $\mathbb{P}(\emptyset) = 0$;
- $\mathbb{P}(\Omega) = 1$;
- Pour toute suite de sous-ensembles $(A_i)_{i \in \mathbb{N}}$ de \mathcal{F} deux à deux disjoints, on a

$$\mathbb{P}\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} \mathbb{P}(A_i).$$

On peut comprendre $\mathbb{P}(A)$ comme la probabilité que le monde réel soit dans l'ensemble A . Ainsi, $\mathbb{P}(\emptyset) = 0$ car il est absolument certain que le monde réel ne soit pas dans l'ensemble vide ! De même, comme Ω est l'ensemble de tous les mondes possibles, monde réel inclus, le monde réel appartient à Ω et donc $\mathbb{P}(\Omega) = 1$. On peut aussi imaginer que Ω est une surface d'aire 1. Alors $\mathbb{P}(A)$ est l'aire de A . Le dernier point signifie que l'aire d'une union disjointe de surfaces est la somme des aires des surfaces.

La proposition qui suit continue des propriétés naturelles sur la mesure \mathbb{P} . Leurs démonstrations sont des exercices ennuyeux (autrement dit, des exercices pour vous) !

Proposition 8. 1. Pour toute suite de sous-ensembles $(A_i)_{i=1..n}$ de \mathcal{F} deux à deux disjoints, on a

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mathbb{P}(A_i).$$

2. Pour tout $A, B \in \mathcal{F}$, $A \cap B = \emptyset$ implique $\mathbb{P}(A \sqcup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
3. Pour tout $A \in \mathcal{F}$, $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$.
4. Pour tout $A, B \in \mathcal{F}$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
5. Pour tout $A, B \in \mathcal{F}$, $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.
6. Pour tout $A, B \in \mathcal{F}$, $\mathbb{P}(A \cap B) \leq \min(\mathbb{P}(A), \mathbb{P}(B))$
7. Pour tout $A, B \in \mathcal{F}$, $\mathbb{P}(A \cup B) \geq \max(\mathbb{P}(A), \mathbb{P}(B))$
8. Pour toute suite de sous-ensembles $(A_i)_{i \in \mathbb{N}}$ de \mathcal{F} , on a

$$\mathbb{P}\left(\bigcup_{i \in \mathbb{N}} A_i\right) \leq \sum_{i \in \mathbb{N}} \mathbb{P}(A_i).$$

9. Pour toute suite de sous-ensembles $(A_i)_{i \in \mathbb{N}}$ de \mathcal{F} avec $A_i \subseteq A_{i+1}$,
alors $\mathbb{P}(A_i)$ est croissante et converge vers $\mathbb{P}(\bigcup_{i \in \mathbb{N}} A_i)$.
10. Pour toute suite de sous-ensembles $(A_i)_{i \in \mathbb{N}}$ de \mathcal{F} avec $A_{i+1} \subseteq A_i$,
alors $\mathbb{P}(A_i)$ est décroissante et converge vers $\mathbb{P}(\bigcap_{i \in \mathbb{N}} A_i)$.

Corollaire 9. Si A_1, A_2, \dots sont des événements presque sûrs alors $\bigcap_{i=1}^{\infty} A_i$ est presque sûre.

DÉMONSTRATION.

$$\begin{aligned} \mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i\right) &= 1 - \mathbb{P}\left(\overline{\bigcap_{i=1}^{\infty} A_i}\right) \\ &= 1 - \mathbb{P}\left(\bigcup_{i=1}^{\infty} \overline{A_i}\right) \\ &\geq 1 - \sum_{i=1}^{\infty} \mathbb{P}(\overline{A_i}) \\ &\geq 1 - 0 = 1 \end{aligned}$$

■

Proposition 10 (formule des probabilités totales). Soit $\{B_i \mid i \in I\}$ une partition de finie ou dénombrable de Ω . Alors :

$$\mathbb{P}(A) = \sum_{i \in I} \mathbb{P}(A \cap B_i).$$

Notations logiques

On utilise parfois les notations des opérateurs de la logique propositionnelle :

- $A \wedge B$, ou “ A, B ” pour $A \cap B$; (A et B)
- $A \vee B$ pour $A \cup B$; (A ou B)
- $\neg A$ pour \overline{A} . (non A)

Exemple 11. $\mathbb{P}(A) = \mathbb{P}(A \wedge B) + \mathbb{P}(A \wedge \neg B)$.

2.6 Espace probabilisé

Un espace probabilisé est la donnée d’un univers Ω , d’une tribu contenant exactement les événements, et d’une mesure de probabilité.

Définition 12. Un **espace probabilisé** est $(\Omega, \mathcal{F}, \mathbb{P})$ est la donnée :

- d’un espace Ω ;
- d’une tribu \mathcal{F} sur Ω ;
- d’une mesure de probabilité $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$.

Exemple 13 (résultat d’un dé).

- $\Omega = \{1, 2, 3, 4, 5, 6\}$;
- $\mathcal{F} = 2^{\{1, 2, 3, 4, 5, 6\}}$;
- Pour tout $A \subseteq \Omega$, $\mathbb{P}(A) = \frac{|A|}{6}$.

Exemple 14 (probabilité subjective sur l’incertitude totale sur un chiffre).

- $\Omega = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$;
- $\mathcal{F} = 2^{\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}}$;
- Pour tout $A \subseteq \Omega$, $\mathbb{P}(A) = \frac{|A|}{10}$.

Exemple 15 (probabilité subjective sur l’incertitude entre un 4 et un 7).

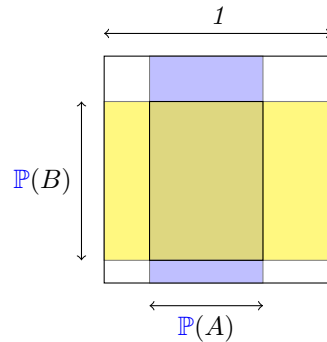


- $\Omega = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$;
- $\mathcal{F} = 2^{\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}}$;
- Pour tout $A \subseteq \Omega$, $\mathbb{P}(A) = 0.8 \times 1_{4 \in A} + 0.2 \times 1_{7 \in A}$.

2.7 Indépendance

Définition 16. Deux événements A et B sont **indépendants** si $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$.

Exemple 17. L'image mentale de deux événements indépendants est de considérer A et B comme des cylindres, dont les génératrices respectives suivent des axes différents.



Notation 18. $A \perp B$.

Définition 19. Des événements indépendants $(A_i)_{i \in I}$ sont **mutuellement indépendants** si pour tout $J \subseteq I$, J finie, $\mathbb{P}(\bigcap_{j \in J} A_j) = \prod_{j \in J} \mathbb{P}(A_j)$.

Attention : indépendance mutuelle \neq indépendance deux à deux

La définition 19 est la définition de l'indépendance mutuelle. Cela implique que les événements soient indépendants deux à deux, c'est-à-dire que pour tout $i, j \in I$, A_i et A_j soient indépendants.

Mais la réciproque n'est pas toujours vraie. Considérons deux lancers de pièces. Considérons les événements :

- A = pile au premier lancer ;
- B = pile au deuxième lancer ;
- C = les résultats des deux lancers sont différents.

On a $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = 1/2$. Et $\mathbb{P}(A \cap B) = \mathbb{P}(A \cap C) = \mathbb{P}(B \cap C) = 1/4$. Donc A , B et C sont 2 à 2 indépendants.

Mais $\mathbb{P}(A \cap B \cap C) = 0 \neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$ donc ils ne sont pas mutuellement indépendants.

Exercices

Donner une définition plus succincte d'une tribu.

Donner un exemple pour Ω pour une suite infini de lancers de dé. Quel est la tribu utilisée ? La mesure de probabilité ?

Montrer que la tribu borélienne est engendrée par l'ensemble des intervalles de la forme $] - \infty, a]$ où $a \in \mathbb{Q}$.

Montrer la proposition 8.

Aller plus loin

Logique probabiliste

Théorie des possibilités

Chapitre 3

Probabilités conditionnelles

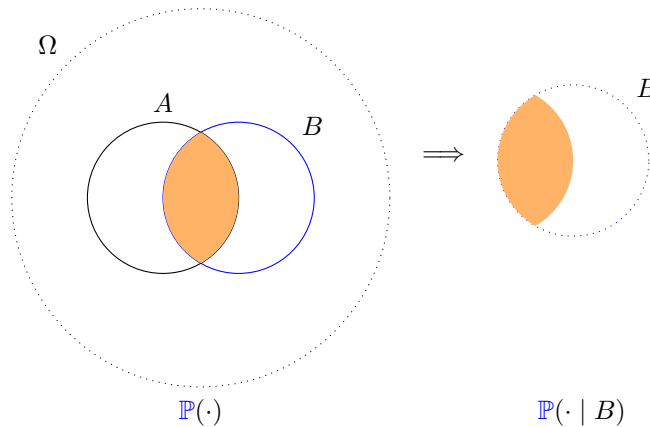
La probabilité conditionnelle correspond au calcul de nouvelles probabilités, en supposant que l'on apprenne qu'un événement B s'est réalisé ou va se réaliser.

3.1 Définition

On suppose que l'événement B est compatible avec notre ensemble de mondes possibles. Dit autrement, l'événement B est probable, c'est-à-dire que $\mathbb{P}(B) > 0$.

Définition 20 (probabilité conditionnelle). Si $\mathbb{P}(B) > 0$, on note $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.

Comme le montre le schéma ci-dessous, cela revient à considérer que le nouvel ensemble Ω est maintenant B , et de normaliser la mesure de probabilité par $\mathbb{P}(B)$.



Exemple 21 (paradoxe des deux enfants). Un parent a deux enfants. L'un d'eux est un garçon. Quel est la probabilité que l'autre enfant soit aussi un garçon ?

$$\mathbb{P}(GG \mid G \cdot \text{ou} \cdot G) = \frac{\mathbb{P}(GG \cap G \cdot \text{ou} \cdot G)}{\mathbb{P}(G \cdot \text{ou} \cdot G)} = \frac{\mathbb{P}(GG)}{\mathbb{P}(G \cdot \text{ou} \cdot G)} = \frac{1/4}{3/4} = 1/3.$$

Mise à jour et révision

La condition B correspond à apprendre que B est vraie. Apprendre un fait est étudié dans plein de sous-domaine de l'intelligence artificielle. En logique, cela revient à ajouter une formule logique supplémentaire à une collection de formules d'une base de connaissance. En raisonnement sur les connaissances, la logique des annonces publiques modélise exactement le fait que les agents d'un système apprenne une formule logique de manière publique. Ici on suppose que B est compatible avec nos croyances. Cela est donné par $\mathbb{P}(B) > 0$. Le cas où B n'est pas compatible avec nos croyances a été étudié en **théorie de la révision de croyance**.

Proposition 22. Si $\mathbb{P}(B) > 0$, alors $(\Omega, \mathcal{F}, \mathbb{P}(\cdot \mid B))$ est encore un espace probabilisé.

Attention, attention !

La notation $\mathbb{P}(A | B)$ n'est pas heureuse car $A | B$ ne désigne pas un événement ! On aurait pu écrire plutôt $\mathbb{P}_B(A)$, mais on préfère utiliser la notation standard $\mathbb{P}(A | B)$ car elle se lit plus facilement.

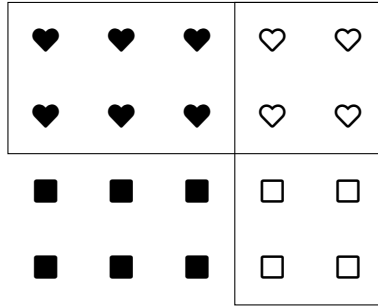
Proposition 23. Si $A \perp B$, alors $\mathbb{P}(A | B) = \mathbb{P}(A)$.

Proposition 24 (règle du produit). Si $\mathbb{P}(A), \mathbb{P}(B) > 0$,

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

Exemple 25. On donne ici deux façons de calculer l'aire de la zone 'cœur et blanc'. D'une part, on considère $\mathbb{P}(\text{cœur} | \text{blanc})$ qui est la proportion de cœurs parmi les objets blancs, multipliée par l'aire des objets blancs. D'autre part, c'est aussi $\mathbb{P}(\text{blanc} | \text{cœur})\mathbb{P}(\text{cœur})$, la proportion de cœurs blancs parmi les cœurs, multipliée par l'aire des cœurs.

$$\begin{aligned} \mathbb{P}(\text{cœur et blanc}) &= \mathbb{P}(\text{cœur} | \text{blanc}) \times \mathbb{P}(\text{blanc}) = \mathbb{P}(\text{blanc} | \text{cœur})\mathbb{P}(\text{cœur}) \\ \frac{4}{20} &= \frac{\frac{4}{8} \times \frac{8}{20}}{\frac{4}{10} \times \frac{10}{20}} \end{aligned}$$




Proposition 26 (Formule des probabilités totales). Si $\Omega = \bigsqcup_{i \in I} E_i$ est une partition finie ou dénombrable. Alors pour tout événement A ,

$$\mathbb{P}(A) = \sum_{i \in I} \mathbb{P}(A | E_i) \times \mathbb{P}(E_i).$$

3.2 Formule de Bayes

Théorème 27 (formule de Bayes). Si $\mathbb{P}(A) > 0$ et $\mathbb{P}(B) > 0$,

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

 **Exemple 28** (problème de Monty-Hall). Bienvenue au jeu télévisé *Let's Make a Deal*. L'animateur Monty Hall vous présente trois portes fermées.



Un cadeau se trouve derrière l'une des portes. Les trois possibilités équiprobables sont :



Le protocole est :

1. On choisit une porte (mettons la porte n° 1)
2. L'animateur nous dévoile une porte vide (disons la porte n° 3) parmi les non choisies

3. L'animateur nous demande si on veut changer de porte ? (i.e. ouvrir finalement la porte n° 2)

$$\begin{aligned}\mathbb{P}(\text{cadeau en 1} \mid \text{porte dévoilée vide} = n^\circ 3) &= \frac{\mathbb{P}(\text{cadeau en 1})}{\mathbb{P}(\text{porte dévoilée vide} = n^\circ 3)} \times \mathbb{P}(\text{porte dévoilée vide} = n^\circ 3 \mid \text{cadeau en 1}) \\ &= \frac{1/3}{1/2} \times 1/2 = 1/3\end{aligned}$$

$$\begin{aligned}\mathbb{P}(\text{cadeau en 2} \mid \text{porte dévoilée vide} = n^\circ 3) &= \frac{\mathbb{P}(\text{cadeau en 2})}{\mathbb{P}(\text{porte dévoilée vide} = n^\circ 3)} \times \mathbb{P}(\text{porte dévoilée vide} = n^\circ 3 \mid \text{cadeau en 2}) \\ &= \frac{1/3}{1/2} \times 1 = 2/3\end{aligned}$$

Exercice 3.1. Que pensez-vous de



$$\begin{aligned}\mathbb{P}(\text{cadeau en 1} \mid \text{rien en 3}) &= \frac{\mathbb{P}(\text{cadeau en 1})}{\mathbb{P}(\text{rien en 3})} \times \mathbb{P}(\text{rien en 3} \mid \text{cadeau en 1}) \\ &= \frac{1/3}{2/3} \times 1 = 1/2?\end{aligned}$$

Exemple 29 (ne tuons pas de personnes inutilement). *Considérons une population de 10000 personnes avec 10 malades sur 10000. Il y a un test de dépistage qui fonctionne à 90% : $\mathbb{P}(\text{positif} \mid \text{malade}) = 0.9$ et $\mathbb{P}(\text{néгатif} \mid \text{non malade}) = 0.9$. Il existe un traitement qui fonctionne à 50% : les autres 50% meurent du traitement. L'État décide de tester tout le monde une fois. Faut-il utiliser le traitement sur les personnes positives ?*



La probabilité qu'une personne au pif soit positive au test vaut :

$$\begin{aligned}\mathbb{P}(\text{positif}) &= \mathbb{P}(\text{positif} \mid \text{malade})\mathbb{P}(\text{malade}) + \mathbb{P}(\text{positif} \mid \text{non malade})\mathbb{P}(\text{non malade}) \\ &= 0.9 \times 1/1000 + 0.1 \times 999/1000 = 0.1008\end{aligned}$$

On calcule la probabilité d'être malade sachant que le test est positif :

$$\begin{aligned}\mathbb{P}(\text{malade} \mid \text{positif}) &= \frac{\mathbb{P}(\text{malade})}{\mathbb{P}(\text{positif})} \times \mathbb{P}(\text{positif} \mid \text{malade}) \\ &= \frac{1/1000}{0.1008} \times 0.9 = 0.0089...\end{aligned}$$

On a :

$$\mathbb{P}(\text{non malade} \mid \text{positif}) = 1 - \mathbb{P}(\text{malade} \mid \text{positif}) = 0.9911....$$

La proportion de personnes non malades et positives est :

$$\mathbb{P}(\text{non malade et positif}) = \mathbb{P}(\text{non malade} \mid \text{positif}) \times \mathbb{P}(\text{positif}) = 0.1$$

et la moitié de ces personnes vont mourir... pour rien ! Sur 10000 personnes, 500 personnes vont vraiment juste mourir pour rien. C'est beaucoup plus que les 10 personnes malades.

Exemple 30 (le procès de Sally Clark, deux bébés morts). Une femme a deux fois de suite un bébé mort. Comme la probabilité d'avoir deux bébés morts est très faible, elle est jugée coupable d'avoir tué ses bébés. Qu'en pensez-vous ?

Raisonnement fait lors du procès

$$\mathbb{P}(1 \text{ enfants mort}) \approx \frac{1}{8500}$$

En supposant que c'est indépendant :

$$\mathbb{P}(2 \text{ enfants morts}) \approx \frac{1}{8500} \frac{1}{8500} \approx \frac{1}{73000000}$$

Comme $\mathbb{P}(2 \text{ enfants morts})$ est petit, elle est coupable.

Explications

Le raisonnement fait lors du procès est fallacieux car on confond $\mathbb{P}(2 \text{ enfants morts})$ petit et $\mathbb{P}(\text{innocent})$ petit. Plus précisément, on a :

$$\mathbb{P}(2 \text{ enfants morts} \mid \text{innocent}) = \frac{1}{73000000}$$

Mais ce qu'il faut calculer c'est $\mathbb{P}(\text{innocent} \mid 2 \text{ enfants morts})$. On a :

$$\mathbb{P}(\text{innocent} \mid 2 \text{ enfants morts}) = \frac{\mathbb{P}(2 \text{ enfants morts} \mid \text{innocent}) \times \mathbb{P}(\text{innocent})}{\mathbb{P}(2 \text{ enfants morts})}$$

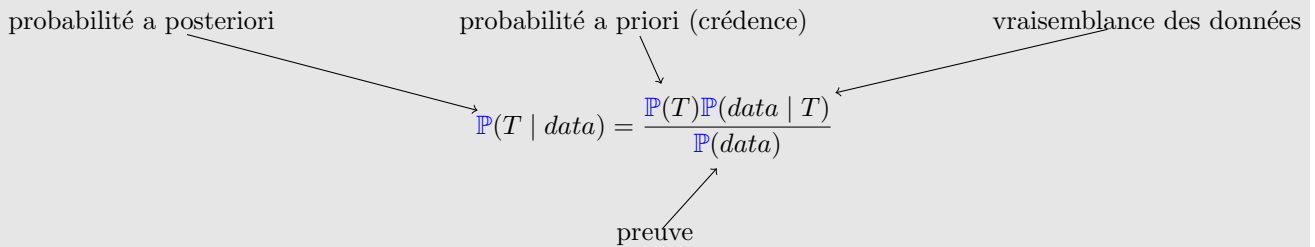
$$\text{On a : } \mathbb{P}(2 \text{ enfants morts}) \approx \frac{1}{73000000}.$$

$\mathbb{P}(\text{innocent}) \approx 1$ car la plupart des mamans ne tuent pas leurs enfants !

On devrait avoir donc $\mathbb{P}(\text{innocent} \mid 2 \text{ enfants morts})$ grand devant $\mathbb{P}(2 \text{ enfants morts} \mid \text{innocent})$.

Inférence bayésienne

On dispose d'une collection de théories T potentielles pour expliquer le monde, autrement dit on dispose de $\mathbb{P}(\cdot \mid T)$. Étant donné une observation $data$, on peut calculer la probabilité que la théorie T soit vraie sachant les données $data$.



$\mathbb{P}(data)$ se calcule via $\mathbb{P}(data) = \sum_{\text{théorie } T'} \mathbb{P}(data \mid T')\mathbb{P}(T')$.

3.3 Indépendance conditionnelle

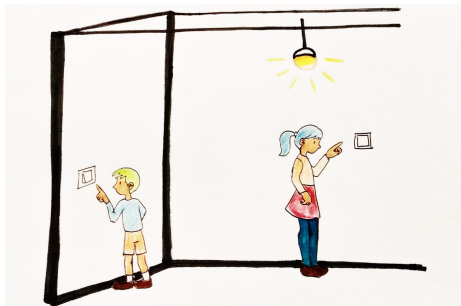
Définition 31. Deux événements A et B sont **indépendants conditionnellement** à C si

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

Notation 32. $A \perp\!\!\!\perp B \mid C$.

Autrement dit, deux événements sont indépendants conditionnellement à C , s'ils sont indépendants après avoir appris C . Comme le montre l'exemple suivant, deux événements indépendants peuvent, en apprenant C ne plus l'être conditionnellement à C .

Exemple 33. On est dans une chambre avec deux interrupteurs pour allumer la lumière. La lumière est allumée si les deux interrupteurs sont dans la même position (tous les deux en haut, ou tous les deux en bas). Deux enfants indépendants s'occupent chacun de leur interrupteurs.



interrupteur 1 en haut $\perp\!\!\!\perp$ interrupteur 2 en haut

interrupteur 1 en haut $\perp\!\!\!\perp$ lumière

interrupteur 2 en haut $\perp\!\!\!\perp$ lumière

Mais interrupteur 1 en haut $\not\perp\!\!\!\perp$ interrupteur 2 en haut \mid lumière.

Exercices

Exercice 3.2. On vous présente n prix dans un certain ordre. A chaque prix, vous décidez soit de l'accepter, soit de le rejeter et de passer au prix suivant. Vous gagnez si vous avez accepté le meilleur prix.

A chaque prix présenté, la seule information que vous avez à votre disposition est la liste des prix déjà vus.

Soit $k \in \mathbb{N}$. Nous allons considérer la stratégie suivante :

— rejeter les k premiers prix ;

— puis accepter le prix qui est meilleur que les k premiers prix

On suppose que les $n!$ ordres des n prix sont équiprobables. L'objectif est de calculer la probabilité de gagner.

1. On note X l'indice du meilleur prix. Que vaut $\mathbb{P}(X = i)$?
2. Écrire la formule des probabilités totales (cf. proposition 26) pour $\mathbb{P}(\text{gagner})$ en considérant les événements $X = 1, X = 2, \dots, X = n$.
3. Que vaut $\mathbb{P}(\text{gagner} \mid X = i)$? (étudier les cas $i \leq k$ et $i > k$ peut aider)
4. Montrer que $\mathbb{P}(\text{gagner})$ est équivalent à $\frac{k}{n} \log\left(\frac{n}{k}\right)$ quand n tend vers l'infini.
5. Nous allons maintenant calculer la valeur de k qui donne la plus grande probabilité de $\mathbb{P}(\text{gagner})$. Pour cela, étudier la fonction $f(x) = \frac{x}{n} \log\left(\frac{n}{x}\right)$.

Chapitre 4

Variables aléatoires et lois

4.1 Variable aléatoire

Définition 34 (variable aléatoire). Une variable aléatoire X réelle est une application¹ de $X : \Omega \rightarrow \mathbb{R}$.

Exemple 35. Soit $\Omega = \{\square\odot, \square\odot, \square\odot, \square\odot, \square\odot, \square\odot, \square\odot, \square\odot, \square\odot, \square\odot, \square\odot, \square\odot\}$.

La variable X donne la valeur du dé :

$X :$	Ω	\rightarrow	\mathbb{R}
		\mapsto	1
		\mapsto	2
		\mapsto	3
		\mapsto	4
		\mapsto	5
		\mapsto	6
		\mapsto	1
		\mapsto	2
		\mapsto	3
		\mapsto	4
		\mapsto	5
		\mapsto	6

Définition 36 (loi d'une variable aléatoire). La loi de X , notée \mathbb{P}_X , est définie pour tout borélien A :

$$\mathbb{P}_X(A) := \mathbb{P}(X \in A) := \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in A\}).$$

Exemple 37. $\mathbb{P}(X \in \{1, 2\}) = \mathbb{P}(\square\odot, \square\odot, \square\odot, \square\odot)$.

Définition 38 (notations plus légères).

$$\mathbb{P}(X = a) := \mathbb{P}_X(\{a\})$$

$$\mathbb{P}(X \leq a) := \mathbb{P}_X(\{x \leq a\})$$

4.2 Indépendance

4.2.1 Deux variables

Définition 39 (deux variables indépendantes). X_1 et X_2 sont indépendantes si pour tout borélien B_1, B_2 , les événements $X_1 \in B_1$ et $X_2 \in B_2$ sont indépendants.

Proposition 40. Soit X_1, X_2 deux variables aléatoires réelles. Sont équivalents :

1. X_1 et X_2 sont indépendantes;
2. Pour tout $t_1, t_2 \in \mathbb{R}$,

$$\mathbb{P}(X_1 \leq t_1 \text{ et } X_2 \leq t_2) = \mathbb{P}(X_1 \leq t_1) \times \mathbb{P}(X_2 \leq t_2).$$

1. mesurable de (Ω, \mathcal{F}) dans $(\mathbb{R}, \text{Borel}(\mathbb{R}))$ où $\text{Borel}(\mathbb{R})$ est la tribu des boréliens. Plus précisément, pour tout $B \subseteq \text{Borel}(\mathbb{R})$, $X^{-1}(B)$ est mesurable. Cela signifie que l'on va attribuer une mesure de probabilité à l'événement $X \in]a, b[$!

Exemple 41 (variables dépendantes). $X_1 =$ résultat d'un dé à 6 faces. $X_2 = X_1 + 1$. Les événements $X_1 = 1$ et $X_2 = 2$ ne sont pas indépendants. En effet,

$$\underbrace{\mathbb{P}(X_1 = 1 \text{ et } X_2 = 3)}_0 \neq \underbrace{\mathbb{P}(X_1 = 1)}_{1/6} \times \underbrace{\mathbb{P}(X_2 = 3)}_{1/6}.$$

Exemple 42 (variables indépendantes). Voici un exemple de deux variables indépendantes :

$X_1 =$ résultat d'un lancer de dé à 6 faces

$X_2 =$ résultat d'un deuxième lancers de dé à 6 faces.

Lemme 43 (lemme des coalitions). Soit X, Y deux variables aléatoires. Et soit $f, g : \mathbb{R} \rightarrow \mathbb{R}$ des fonctions². Alors :

X et Y indépendantes implique $f(X)$ et $g(Y)$ indépendantes.

DÉMONSTRATION.

$$\begin{aligned} \mathbb{P}(f(X) \in A \text{ et } g(Y) \in B) &= \mathbb{P}(X \in f^{-1}(A) \text{ et } Y \in g^{-1}(B)) \\ &= \mathbb{P}(X \in f^{-1}(A)) \times \mathbb{P}(Y \in g^{-1}(B)) \text{ car } X \text{ et } Y \text{ sont indépendantes} \\ &= \mathbb{P}(f(X) \in A) \times \mathbb{P}(g(Y) \in B) \end{aligned}$$

■

4.2.2 Plusieurs variables

Définition 44 (variables indépendantes). Soit $(X_i)_{i \in I}$ une collection finie ou dénombrable de variables aléatoires. On dit que les variables $(X_i)_{i \in I}$ sont **mutuellement indépendantes** si les événements $X_i \in B_i$ pour $i \in I$ sont indépendants (cf. définition 19 page 16).

En appliquant la définition définition 19 page 16, cela donne :

Proposition 45. Les variables $(X_i)_{i \in I}$ sont mutuellement indépendantes ssi pour tout sous-ensemble finie $J \subseteq I$, pour toute collection de boréliens $(B_i)_{i \in J}$ on a :

$$\mathbb{P}\left(\bigcap_{i \in J} X_i \in B_i\right) = \prod_{i \in J} \mathbb{P}(X_i \in B_i)$$

Définition 46. X_1, \dots, X_n sont **iid** (indépendantes identiquement distribuées) si elles sont mutuellement indépendantes et de même loi.

Exemple 47. $X_i =$ résultats du i -ème lancer d'un dé à 6 faces. Les lancers sont indépendants.

4.3 Lois discrètes

Définition 48. Une va X est **discrète** s'il existe un ensemble D fini ou dénombrable tel que $\mathbb{P}(X \in D) = 1$.

Définition 49. D s'appelle le **support** de X .

4.3.1 Dirac

C'est la distribution la plus simple qui correspond à ne pas avoir d'aléatoire.

Définition 50. X suit une **loi de Dirac** en a si $\mathbb{P}(X = a) = 1$. La loi de Dirac en a est notée δ_a .

Dirac

Dirac est une personne. C'est le mathématicien Paul Dirac. Mais souvent on dit un dirac pour dire une distribution de Dirac, ou loi de Dirac.

2. ... mesurables

4.3.2 Loi de Bernoulli

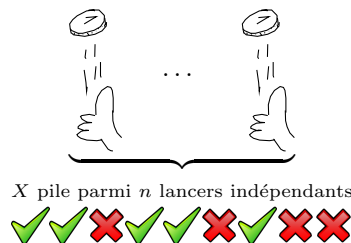
Une loi de Bernoulli $\mathcal{B}(p)$ modélise le lancer d'une pièce de monnaie pipé où il y a une probabilité p de faire pile, et $1 - p$ de faire face. Pile est comptabilisé comme 1, et face comme 0.

Définition 51 (loi de Bernoulli). $\mathcal{B}(p)$ est la loi d'une variable aléatoire X avec
$$\begin{cases} \mathbb{P}(X = 1) = p & \checkmark \text{ succès} \\ \mathbb{P}(X = 0) = (1 - p) & \times \text{ échec} \end{cases}$$

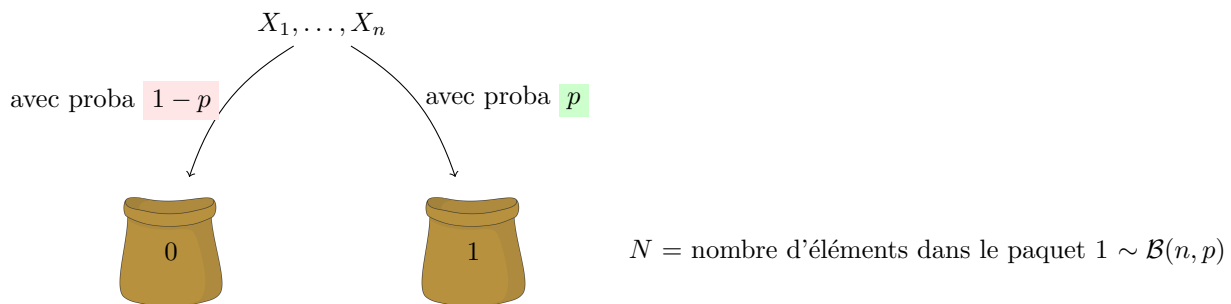
4.3.3 Loi binomiale

Une loi binomiale $\mathcal{B}(n, p)$ correspond au **nombre de succès sur n essais iid**. Dit autrement, c'est le nombre de fois que l'on a obtenu pile avec une pièce de monnaie, lancée n fois, et toujours avec probabilité p de faire pile sur un lancer. Les lancers sont bien indépendants. Hors de question, par exemple que pour le 5e lancer, on recopie la valeur obtenu au 4e lancer ; les 4e et 5e lancers sont indépendants, c'est-à-dire qu'il faut bien relancer la pièce. Aussi, le résultat du 4e lancer n'influence pas du tout le résultat du 5e lancer. On voit que $\mathcal{B}(1, p) = \mathcal{B}(p)$.

Définition 52 (loi binomiale). $\mathcal{B}(n, p)$ est la loi de $N = X_1 + \dots + X_n$ où les $X_i \sim \mathcal{B}(p)$ sont iid.



Dit autrement, la loi binomiale compte le nombre d'éléments qui sont dans le sac 1 si la probabilité de mettre un élément vaut p (loi de Bernoulli de paramètre p), et si on a n éléments en tout.



Proposition 53. Si $N \sim \mathcal{B}(n, p)$, alors pour tout $k \in \{0, \dots, n\}$, on a $\mathbb{P}(N = k) = \binom{n}{k} \times p^k \times (1 - p)^{n-k}$.

Dénombrement et loi binomiale

$p^k(1 - p)^{n-k}$ est la probabilité d'avoir (par exemple) d'abord k piles, puis ensuite les $n - k$ faces. Mais la position des piles est tout au début.

$\binom{n}{k}$ est le nombre de sous-ensemble de cardinal k , i.e. le nombre de façons de placer ces k piles. D'où $\binom{n}{k} p^k (1 - p)^{n-k}$ qui est la probabilité d'avoir k piles parmi n lancers.

4.3.4 Loi géométrique

La loi géométrique, aussi appelée loi de Pascal, correspond au **temps d'attente jusqu'à un succès** (essai avec succès compris). Par exemple, le nombre de lancers de pièce de monnaie jusqu'à avoir pile.



Définition 54 (loi géométrique). $\mathcal{G}(p)$ est la loi de $T = \min(k \in \mathbb{N}^* \mid X_k = 1)$ où $X_1, X_2, \dots \sim \mathcal{B}(p)$ iid.

Proposition 55. Si $T \sim \mathcal{G}(p)$, pour tout $k \in \mathbb{N}^*$, $\mathbb{P}(T = k) = (1 - p)^{k-1} p$.

Proposition 56. $\mathbb{P}(T \geq k) = (1 - p)^{k-1}$.

DÉMONSTRATION.

L'événement $T \geq k$ consiste avoir $k - 1$ échecs au début puis n'importe quoi :

$$\underbrace{\text{XXXXXX}}_{k-1 \text{ échecs}} \text{?????} \dots$$

■

Proposition 57 (sans mémoire). Soit $T \sim \mathcal{G}(p)$. Soit $s, t \in \mathbb{N}$. $\mathbb{P}(T \geq s + t \mid T \geq s) = \mathbb{P}(T \geq t)$.

4.3.5 Lois de Poisson

Une loi de Poisson est l'**analogue d'une loi binomiale** mais pour un **temps continu** et pour des **événements rares**. Pour une loi binomiale, les instants sont $\{1, \dots, n\}$ et il y a une probabilité p d'avoir un succès à chaque instant $i \in \{1, \dots, n\}$. Pour une loi de Poisson, imaginons un intervalle de temps $[0, \tau]$. Découpons le en intervalles infinitésimaux de longueur dt . Autrement dit :

$$n = \frac{\tau}{dt}.$$

Ces intervalles sont si petits qu'il n'y a qu'un seul lancer de pièce sur chacun d'eux. La probabilité d'avoir un succès est très petite :

$$\mathbb{P}(\text{succès sur un intervalle de longueur } dt) = \nu dt = p$$

où ν est un nombre positif que l'on appelle intensité du processus, c'est le nombre moyen de succès par unité de temps. La loi de Poisson est la loi de

$$N = \text{nombre de succès sur l'intervalle } [0, \tau].$$

Elle est caractérisée par le paramètre $\lambda = \nu \times \tau = pn$. Le paramètre λ est le nombre de succès moyen sur l'intervalle $[0, \tau]$. Dit autrement, une loi de Poisson est une approximation de la loi binomiale quand p est très petit, et n grand, et pn est de l'ordre de grandeur de $\lambda \in \mathbb{R}$. C'est le **nombre de succès pour des événements rares**.

Lois des événements rares

Si p est très petit, et n grand, et pn est de l'ordre de grandeur de $\lambda \in \mathbb{R}$, on approxime la loi binomiale par une loi de Poisson :

$$\mathbb{P}(N = k) = \binom{n}{k} p^k (1 - p)^{n-k} = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \left(1 - \frac{\lambda}{n}\right)^n \times \frac{\lambda^k}{k!} \frac{n! (1 - \frac{\lambda}{n})^{-k}}{(n-k)! n^k} \xrightarrow{n \rightarrow +\infty} \frac{\lambda^k}{k!} e^{-\lambda}$$

L'énoncé formel de la loi des événements rares est donné en théorème 165 page 56.

Définition 58 (loi de Poisson). $\mathcal{P}(\lambda)$ est donnée par $\mathbb{P}(N = k) = \frac{\lambda^k}{k!} e^{-\lambda}$.

Exemple 59. $N = \text{nombre d'arrivants à un guichet par minutes}$.

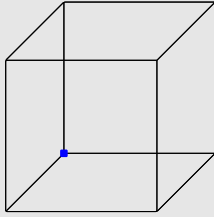
Exercice 4.1. Un industriel fabrique des pièces métalliques. La probabilité qu'une pièce soit défectueuse est de $p = 0.005$. Là, on va contrôler 800 pièces.

1. A priori, quel est la probabilité que plus de 6 pièces soient défectueuses ?
2. Quel est la probabilité qu'au moins deux pièces soient défectueuses ?
3. Probabilité qu'exactement 3 pièces soient défectueuses ?
4. Comparer les résultats de la modélisation avec une loi binomiale, et avec la loi de Poisson

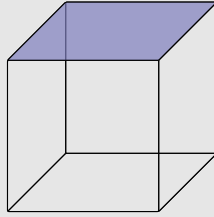
4.4 Lois continues (à densité)

Discret VS continu

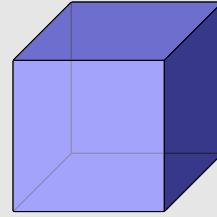
On rencontre ce phénomène en électricité. Une charge électrique peut être entièrement confinée en un seul point (charge discrète), ou alors étalée à la surface d'un objet (densité).



charge discrète en 1 point



charge étalée sur une face



charge étalée dans le volume du cube

Définition 60. Une **densité de probabilité** est une fonction $p : \mathbb{R} \rightarrow \mathbb{R}^+$ telle que $\int_{-\infty}^{+\infty} p(x)dx = 1$.

Définition 61. Une variable réelle X est continue de densité p si $\mathbb{P}(X \leq t) = \int_{-\infty}^t p(x)dx$.

Proposition 62. Soit X une va à densité. On a :

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a < X < b) = \int_a^b p(x)dx.$$

En particulier, $\mathbb{P}(X = a) = 0$.

4.4.1 Lois uniformes

Définition 63. Soit $a < b$. La **loi uniforme** $\mathcal{U}(a, b)$ est de densité $p(x) = \frac{1}{b-a} \mathbf{1}_{[a, b]}$.



4.4.2 Lois exponentielles

Les lois exponentielles modélisent des **temps d'attente continu jusqu'à un succès** ou de manière équivalente **durée entre deux succès**. Elles sont le pendant continu des lois géométriques, et sont sans mémoire.

Exemple 64.

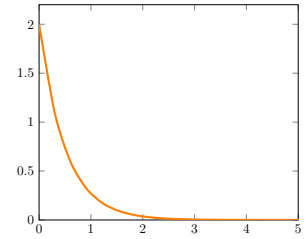
- la durée entre les arrivées de deux clients dans un magasin ;
- la durée entre les arrivées de deux personnes à un arrêt de bus ;
- le temps passé à un guichet ;
- la durée de vie d'un composant électronique qui ne s'use pas, mais qui peut tomber en panne à cause d'un événement extérieur ;
- la durée d'être au chômage.

Soit ν le nombre moyen de succès par une unité de temps. Supposons que l'expérience est sur un intervalle de temps $[0, \tau]$. On divise cette intervalle en n petits morceaux, n très grand. La probabilité d'avoir un succès sur un petit intervalle est de $\nu \frac{\tau}{n}$. Le temps mesuré T' en nombre de petit intervalle suit alors une loi géométrique $\mathcal{G}(\nu \tau n)$. Le temps d'attente est $T = T' \frac{\tau}{n}$. On a :

$$\mathbb{P}(T > s) = \mathbb{P}(T' > s \frac{n}{\tau}) = (1 - \nu \frac{\tau}{n})^{s \frac{n}{\tau}} = \left(1 - \frac{1}{\frac{n}{\nu \tau}}\right)^{\nu s} \xrightarrow{n \rightarrow \infty} e^{-\nu s}$$

On définit la loi exponentielle pour que $\mathbb{P}(T > s) = e^{-\nu s}$.

Définition 65. La **loi exponentielle** $\mathcal{E}(\nu)$ est de densité $p(x) = \nu e^{-\nu x} \mathbf{1}_{\mathbb{R}_+}$.



Proposition 66. $\mathbb{P}(T > s) = e^{-\nu s}$.

DÉMONSTRATION.

On a $\mathbb{P}(T > s) = \int_s^\infty \nu e^{-\nu x} dx = e^{-\nu s}$. ■

L'aspect “sans mémoire” est formalisé par la proposition suivante.

Proposition 67. Soit $T \sim \mathcal{E}(\nu)$. Soit $s, t \in \mathbb{R}$. $\mathbb{P}(T > s + t \mid T > s) = \mathbb{P}(T > t)$.

DÉMONSTRATION.

On a : $\mathbb{P}(T > s + t \mid T > s) = \frac{\mathbb{P}(T > s + t)}{\mathbb{P}(T > s)} = \frac{e^{-\nu(s+t)}}{e^{-\nu s}} = e^{-\nu t} = \mathbb{P}(T > t)$. ■

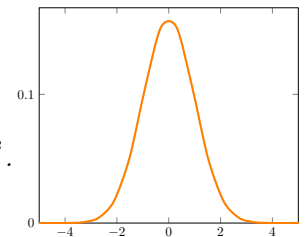
Comparaison entre les différents lois

	Si le temps est ...	
	discret	continu
le nombre de succès dans un intervalle de temps fixe suit une...	loi binomiale $\mathcal{B}(n, p)$	loi de Poisson $\mathcal{P}(\lambda)$
le temps d'attente jusqu'à succès suit une...	loi géométrique $\mathcal{G}(p)$	loi exponentielle $\mathcal{E}(\nu)$

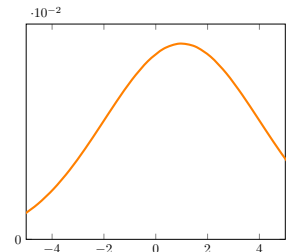
4.4.3 Lois normales

Les lois normales **approximent la somme de variables aléatoires iid** (cf. théorème centrale limite, Théorème 390). Leurs densités ont une forme de **cloche**. C'est assez surprenant car cette approximation fonctionne quelque soit la loi des variables aléatoires ! Mieux, cela fonctionne même pour des lois discrètes !

Définition 68. La **loi normale centrée réduite** $\mathcal{N}(0, 1)$ est de densité $p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$.



Définition 69. Soit $\mu \in \mathbb{R}$. Soit $\sigma > 0$. La **loi normale** $\mathcal{N}(\mu, \sigma^2)$ est de densité :



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

Loi normale

N'ayez pas peur de l'expression de la densité ! Le $\frac{1}{2}$ c'est pour que la variance (voir définition 103 page 36) vaille 1. Le $\frac{1}{\sqrt{2\pi}}$ est juste le facteur de normalisation, car il faut que la l'intégrable de la densité sur $]-\infty, +\infty[$ vaille un.

Proposition 70 (intégrale de Gauss). $\int_{-\infty}^{+\infty} e^{-\frac{1}{2}x^2} dx = \sqrt{2\pi}$.

DÉMONSTRATION.

On pose l'intégrale à calculer : $I = \int_{-\infty}^{+\infty} e^{-\frac{1}{2}x^2} dx$.

On peut calculer le carré de I comme suit :

$$\begin{aligned} I^2 &= \left(\int_{-\infty}^{+\infty} e^{-\frac{1}{2}x^2} dx \right) \left(\int_{-\infty}^{+\infty} e^{-\frac{1}{2}x^2} dx \right) \\ &= \left(\int_{-\infty}^{+\infty} e^{-\frac{1}{2}x^2} dx \right) \left(\int_{-\infty}^{+\infty} e^{-\frac{1}{2}y^2} dy \right) \quad (\text{on renomme juste la variable muette } x \text{ en } y) \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}(x^2+y^2)} dx dy \end{aligned}$$

On reconnaît une intégrale avec $x^2 + y^2$ qui n'est que la distance r au carré du point (x, y) à l'origine. Bref, on va passer en coordonnée polaire pour faire apparaître explicitement cette distance r (le rayon) dans le calcul.

On pose donc :

$$\begin{aligned} x &= r \cos \theta \\ y &= r \sin \theta \end{aligned}$$

L'élément d'aire infinitésimal $dx dy$ est maintenant $r dr d\theta$. Ainsi on a :

$$\begin{aligned} I^2 &= \int_{\theta=0}^{2\pi} \int_{r=0}^{+\infty} e^{-\frac{1}{2}r^2} r dr d\theta \\ &= 2\pi \int_{r=0}^{+\infty} e^{-\frac{1}{2}r^2} r dr \end{aligned}$$

On fait maintenant le changement de variable $t = \frac{r^2}{2}$, $dt = r dr$ et on obtient :

$$\begin{aligned} I^2 &= 2\pi \int_{t=0}^{+\infty} e^{-t} dt \\ &= 2\pi [-e^{-\infty} - -e^0] \\ &= 2\pi \end{aligned}$$

D'où $I = \frac{1}{\sqrt{2\pi}}$. ■

On peut passer d'une loi normale centrée réduite à n'importe quelle loi normale : on multiplie par l'écart-type σ puis on ajoute la moyenne μ . Depuis n'importe quelle loi normale de moyenne μ et d'écart-type σ , on obtient une loi normale centrée réduite en retranchant la moyenne μ , puis en divisant par l'écart-type σ .

Proposition 71 (loi normale centrée réduite \rightsquigarrow n'importe quelle loi normale). Soit $\mu \in \mathbb{R}$ et $\sigma > 0$.

$$\begin{aligned} X \sim \mathcal{N}(0, 1) &\implies \mu + \sigma X \sim \mathcal{N}(\mu, \sigma^2) \\ \frac{Y - \mu}{\sigma} \sim \mathcal{N}(0, 1) &\iff Y \sim \mathcal{N}(\mu, \sigma^2) \end{aligned}$$

4.5 Densité marginale

Définition 72. Soit $X = (X_1, X_2)$ une variable aléatoire sur \mathbb{R}^2 de fonction de densité p_X . Les **densités marginales** sont :

$$p_{X_1}(x_1) = \int_{\mathbb{R}} p_X(x_1, x_2) dx_2,$$

$$p_{X_2}(x_2) = \int_{\mathbb{R}} p_X(x_1, x_2) dx_1.$$

Exemple 73. L'image ci-dessous montre une densité un peu farfelu où il y a une forte probabilité de choisir un point dans la 'zone des deux yeux et de la bouche' ! Les densités marginales sont montrées à droite et en bas.



Définition 74 (densité conditionnelle). Si $p_{X_2}(x_2) > 0$, on définit

$$p_{X_1}(x_1 | x_2) = \frac{p_X(x_1, x_2)}{p_{X_2}(x_2)}.$$

Proposition 75 (théorème de Bayes pour les densités). Si $p_{X_1}(x_1) > 0$, $p_{X_2}(x_2) > 0$ alors

$$p_{X_1}(x_1 | x_2) = \frac{p_{X_1}(x_1)p_{X_2}(x_2 | x_1)}{p_{X_2}(x_2)}$$

DÉMONSTRATION.

On obtient le théorème de Bayes pour les densités à partir des deux définitions suivantes :

$$p_{X_1}(x_1 | x_2) = \frac{p_X(x_1, x_2)}{p_{X_2}(x_2)}.$$

$$p_{X_2}(x_2 | x_1) = \frac{p_X(x_1, x_2)}{p_{X_1}(x_1)}.$$

■

Une densité vérifie les lois de la théorie des probabilités !

La définition des densités marginales correspondent à la formule des probabilités totales (proposition 10 page 15), la définition de la densité conditionnelle à la définition des probabilités conditionnelles (définition 20 page 17)... et le théorème de Bayes s'écrit bien sûr de la même façon !

4.6 Fonction de répartition

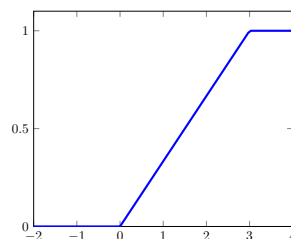
La fonction de répartition d'une variable aléatoire X donne pour chaque réel t , la probabilité que X soit inférieur ou égal à t .

Définition 76 (fonction de répartition). Soit X une va réelle. La fonction de répartition de X est :

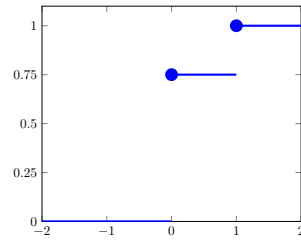
$$F_X : \mathbb{R} \rightarrow [0, 1]$$

$$t \mapsto \mathbb{P}(X \leq t).$$

Exemple 77. (fonction de répartition de la loi uniforme sur $[0, 3]$)



Exemple 78 (fonction de répartition d'une loi de Bernouilli de probabilité 1/4).



Théorème 79. $F_X = F_Y$ ssi X et Y ont même loi.

DÉMONSTRATION.

Si $F_X = F_Y$, cela veut dire que $\mathbb{P}_X(]-\infty, t]) = \mathbb{P}_Y(]-\infty, t])$ pour tout $t \in \mathbb{R}$. Mais comme les $]-\infty, t]$ engendrent tous les boréliens, finito. ■

Proposition 80 (propriétés d'une fonction de répartition).

1. F_X est croissante
2. F_X est continue à droite.
3. $\lim_{t \rightarrow -\infty} F_X(t) = 0$ et $\lim_{t \rightarrow +\infty} F_X(t) = 1$

Proposition 81. Il y a un nombre fini ou dénombrable de points de discontinuités : ce sont les $t \in \mathbb{R}$ avec $\mathbb{P}(X = t) > 0$.

4.7 Loix mélanges (mixture)

Définition 82. Considérons n lois de densités f_1, \dots, f_n et des poids $w_1, \dots, w_n \in [0, 1]$ avec $\sum_i w_i = 1$. La loi de mélange correspondante est décrite par la densité

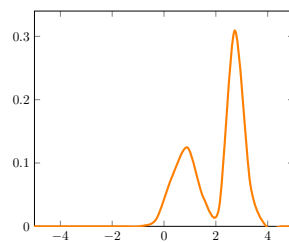
$$x \mapsto \sum_i w_i f_i(x)$$

Exemple 83 (hauteur de personnes). On peut supposer que la hauteur des personnes est modélisée par la loi

$$0.4\mathcal{N}(0.8, 0.5^2) + 0.6\mathcal{N}(2.8, 0.3^2).$$

Il s'agit de la loi dont la densité est

$$p(x) = 0.4 \frac{1}{0.5\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-0.8}{0.5})^2} + 0.6 \frac{1}{0.3\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-2.8}{0.3})^2}$$



Mélange de lois gaussiennes et classification automatique

En apprentissage automatique, le mélange de lois gaussiennes est utilisée pour faire du clustering (cf. section 11.4).

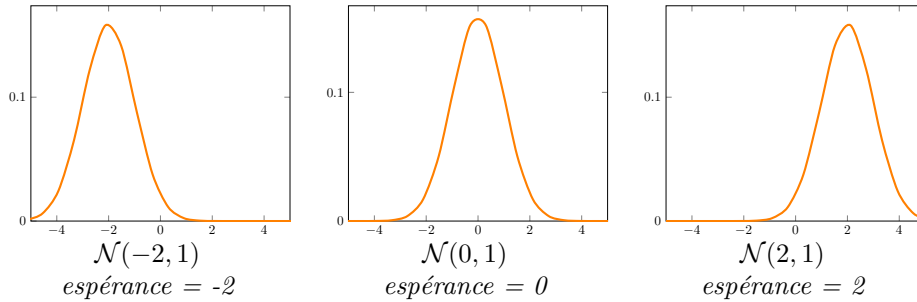
Chapitre 5

Propriétés sur les distributions

5.1 Espérance

L'espérance d'une variable aléatoire X est la moyenne des valeurs que prend la variable. Attention, l'espérance n'est pas forcément une valeur prise par la variable.

Exemple 84 (loi normale de différente espérance).



Définition 85 (espérance finie / intégrable). X est d'espérance finie / intégrable si

$$\int_{\Omega} |X(\omega)| d\mathbb{P}(\omega) < +\infty.$$

Définition 86 (espérance). Si X est d'espérance finie, alors l'espérance de X est

$$\mathbb{E}X := \int_{\Omega} X(\omega) d\mathbb{P}(\omega).$$

Définition 87 (loi centrée). Une va X est centrée si $\mathbb{E}X = 0$.

5.1.1 Théorème de transfert

Le théorème de transfert permet de transférer l'intégration de l'espace abstrait Ω sur l'espace concret \mathbb{R} sur lequel la variable prend ses valeurs. Alors que dans la définition donnée plus haut, la mesure de l'espace abstrait est \mathbb{P} , là la mesure sur l'espace concret est \mathbb{P}_X .

Proposition 88 (théorème de transfert).

$$\mathbb{E}X := \int_{\mathbb{R}} x d\mathbb{P}_X(x).$$

Proposition 89 (théorème de transfert). Soit $\varphi : \mathbb{R} \rightarrow \mathbb{R}$. Alors :

$$\mathbb{E}(\varphi(X)) := \int_{\mathbb{R}} \varphi(x) d\mathbb{P}_X(x).$$

5.1.2 Formules dans le cas discret et continue

Proposition 90. Si X discrète de support D alors la va $\varphi(X)$ est intégrable ssi $\sum_{i \in D} |\varphi(i)| \mathbb{P}(X = i) < +\infty$. On a alors :

$$\mathbb{E}(\varphi(X)) = \sum_{i \in D} \varphi(i) \mathbb{P}(X = i).$$

 **Exemple 91** (espérance de la loi géométrique, voir définition 54 page 25).

$$\begin{aligned} \mathbb{E}(T) &= \sum_{k \in \mathbb{N}^*} k \mathbb{P}(T = k) \\ &= \sum_{k \in \mathbb{N}^*} kp(1-p)^{k-1} \end{aligned}$$

Et là, on est a priori bloqué dans le calcul mais en fait non. On pose la série génératrice :


$$f(x) = \sum_{k=0}^{+\infty} x^k = \frac{1}{1-x}.$$

En dérivant cette série génératrice on obtient :

$$f'(x) = \sum_{k=1}^{+\infty} kx^{k-1} = \frac{1}{(1-x)^2}.$$

Et là, on reconnaît l'expression et on peut remplacer :

$$\begin{aligned} \mathbb{E}(T) &= pf'(1-p) \\ &= \frac{p}{p^2} = \frac{1}{p} \end{aligned}$$

 **Exemple 92** (espérance de la loi de Poisson, voir définition 58 page 26).

On regarde la série génératrice :

$$e^\lambda = \sum_{k=0}^{+\infty} \frac{\lambda^k}{k!}.$$

En dérivant cette série génératrice on obtient :

$$e^\lambda = \lambda \sum_{k=1}^{+\infty} k \frac{\lambda^{k-1}}{(k-1)!}.$$

$$\begin{aligned} \mathbb{E}(N) &= \sum_{k \in \mathbb{N}^*} k e^\lambda \frac{\lambda^k}{k!} \\ &= \lambda e^{-\lambda} e^\lambda \\ &= \lambda. \end{aligned}$$

Intuitivement, λ est le nombre moyen de succès. Donc c'est normal de retrouver ce résultat ici.

Proposition 93. Si X est à densité p alors la va $\varphi(X)$ est intégrable ssi $\int_{\mathbb{R}} |\varphi(x)| p(x) dx < +\infty$. On a alors :

$$\mathbb{E}(\varphi(X)) = \int_{\mathbb{R}} \varphi(x) p(x) dx < +\infty.$$

 **Exemple 94** (espérance de la loi exponentielle, voir définition 65 page 27). Soit $T \sim \mathcal{E}(\nu)$.

$$\begin{aligned}
\mathbb{E}(T) &= \int_0^{+\infty} x\nu e^{-\nu x} dx \\
&= [x(-e^{-\nu x})]_0^{+\infty} - \int_0^{+\infty} e^{-\nu x} dx \\
&= -\left[\frac{1}{\nu}e^{-\nu x}\right]_0^{+\infty} \\
&= \frac{1}{\nu}
\end{aligned}$$

Intuitivement, si ν est le nombre moyen de succès, il est normal d'attendre en moyenne $\frac{1}{\nu}$.

5.1.3 Propriétés

Proposition 95. $\mathbb{E}(\mathbf{1}_A) = \mathbb{P}(A)$

Proposition 96 (linéarité de l'espérance).

$$\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$$

$$\mathbb{E}(aX) = a\mathbb{E}X$$

Théorème 97. X, Y intégrables et **indépendantes** XY intégrable implique $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

DÉMONSTRATION.

Montrons le dans le cas discret (le cas général, il faut le théorème de Fubini et le but premier de ce poly est de comprendre l'essentiel des probas sans avoir trop de théories trop matheuses).

$$\begin{aligned}
\mathbb{E}(XY) &= \sum_{n,k \in \mathbb{N}} nk\mathbb{P}(X = n \text{ et } Y = k) \\
&= \sum_{n,k \in \mathbb{N}} nk\mathbb{P}(X = n)\mathbb{P}(Y = k) \\
&= \sum_{n \in \mathbb{N}} n\mathbb{P}(X = n) \sum_{k \in \mathbb{N}} k\mathbb{P}(Y = k) \\
&= \mathbb{E}X\mathbb{E}Y
\end{aligned}$$

■

Exemple 98 (contre exemple). Prenons X le résultat d'un lancer de pièce. Bien sûr, X et X sont dépendantes. On a :

$$\underbrace{\mathbb{E}(X \times X)}_{0.5} \neq \underbrace{\mathbb{E}(X)\mathbb{E}(X)}_{0.25}.$$

5.1.4 Technique de la fonction test

Bien sûr, deux variables X et Y de même espérance ne suivent pas forcément la même loi. Mais par contre, si on regarde l'espérance de $\varphi(X)$ et $\varphi(Y)$ pour une classe assez grande de fonctions φ qui viennent un peu badigeonner toutes les valeurs réelles possibles, alors on peut conclure que X et Y suivent la même loi. C'est ce que l'on appelle la **technique de la fonction test**.

Théorème 99 (technique de la fonction test). Soit X une variable aléatoire réelle. Soit μ une distribution de probabilité sur \mathbb{R} .

X suit une loi μ ssi pour toute fonction $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ continue et à support compact, $\mathbb{E}(\varphi(X)) = \int_{\mathbb{R}} \varphi(x)d\mu(x)$
ssi pour toute fonction $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ positive bornée, $\mathbb{E}(\varphi(X)) = \int_{\mathbb{R}} \varphi(x)d\mu(x)$

5.2 Variance

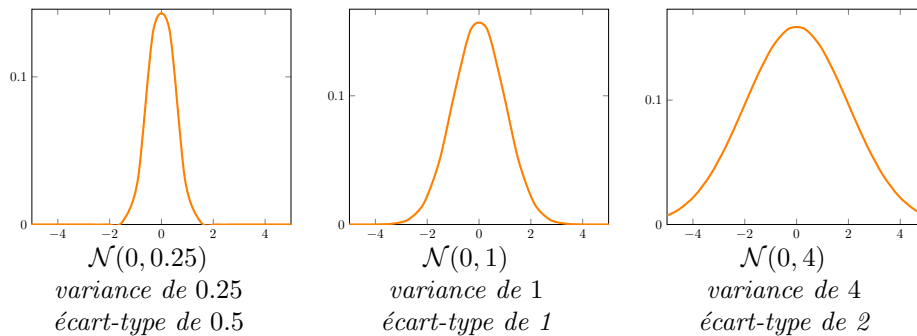
L'objectif est de mesurer l'écart par rapport à la moyenne. Typiquement, une variable aléatoire 'non aléatoire', par exemple une variable qui vaut toujours 42 est de variance nulle car il n'y a absolument jamais d'écart avec la moyenne 42.

La variance est défini comme l'espérance des écarts au carré par rapport à la moyenne. Du coup, la variance de X n'est définie que lorsque X^2 est intégrable.

Mesurer l'écart avec la moyenne

La variance n'est pas le seul moyen de mesurer l'écart avec la moyenne. On peut citer aussi la **déviance absolue** définie par $\mathbb{E}(|X - \mathbb{E}(X)|)$. La principale justification est technique (dérivabilité etc.).

Exemple 100 (lois normales de différentes variances).



Notation 101 (loi normale). $\mathcal{N}(\mu, \sigma^2)$

μ	espérance
σ	écart-type
σ^2	variance

Définition 102 (moment d'ordre 2). X admet un moment d'ordre 2 si X^2 est intégrable, i.e. si $\int |X|^2(\omega) d\mathbb{P}(\omega) < +\infty$.

Définition 103 (variance). Si X admet un moment d'ordre 2, la variance de X est $\mathbb{V}(X) := \mathbb{E}((X - \mathbb{E}X)^2)$.

Définition 104 (écart-type). L'écart-type est $\sqrt{\mathbb{V}(X)}$.

Exemple 105. Soit X une variable de Bernoulli de paramètre p . Son espérance $\mathbb{E}X$ est p . La variance est

$$\begin{aligned}
 \mathbb{V}(X) &= \mathbb{E}((X - p)^2) \\
 &= (0 - p)^2(1 - p) + (1 - p)^2p = p^2(1 - p) + (1 - p)^2p \\
 &= p(1 - p)[p + 1 - p] \\
 &= p(1 - p).
 \end{aligned}$$

Définition 106 (va réduite). Une va X est réduite si $\mathbb{V}(X) = 1$.

5.2.1 Propriétés

Proposition 107 (formule de Koenig-Huygens). $\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$.

Le cas limite est la variance nulle : on obtient alors $\mathcal{N}(0, 0)$ qui est la loi Dirac en 0. Mais en fait, personne ne note $\mathcal{N}(0, 0)$.

Proposition 108. $\mathbb{V}(aX) = a^2\mathbb{V}(X)$.

Corollaire 109. $\frac{X}{\sqrt{\mathbb{V}(X)}}$ est réduite.

Unités : et si X était une distance en mètres

Supposons que la variable X est une distance mesurée en mètres.

Quantités	Unités
X	m
espérance $\mathbb{E}X$	m
écart-type $\sqrt{\mathbb{V}(X)}$	m
variance $\mathbb{V}(X)$	m^2

5.3 Covariance

On cherche à mesurer si deux variables X et Y sont corrélées, c'est-à-dire si un écart de X à sa moyenne, se traduit aussi par un écart de Y à sa moyenne.

Exemple 110 (deux lancers de dé). *Soit X le résultat d'un premier lancer, et Y le résultat d'un deuxième lancer indépendant. A priori, si X vaut 6, il y a autant de chances que Y vaille 6 ou 1. Ça ne change rien. Les variables X et Y sont décorrélées.*

Exemple 111 (taille et poids). *On considère une population d'individus avec des tailles et des poids différents. A priori, plus un individu est grand, plus il est lourd. On dit alors la taille et le poids sont corrélée positivement.*

Corrélation \neq Causalité

Dans une station balnéaire, le nombre de coup de soleil est corrélé positivement au nombre de lunettes de soleil achetées. Pourtant, un coup de soleil ne donne pas envie d'acheter des lunettes de soleil. Et les lunettes de soleil ne cause pas de coup de soleil.

Exemple 112. *On considère 6 lancers de dé.*

X = nombre de \blacksquare obtenus.

Y = nombre de \square obtenus.

On obtient de \blacksquare , moins on obtient de \square : X et Y sont corrélés négativement.

Pour mesurer la corrélation, on définit la covariance comme l'espérance du produit des écarts.

Définition 113 (covariance). *La covariance est $\text{cov}(X, Y) := \mathbb{E}((X - \mathbb{E}X) \times (Y - \mathbb{E}Y))$.*

Proposition 114 (formule de König-Huygens). $\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.

DÉMONSTRATION.

$$\begin{aligned}
 \text{cov}(X, Y) &= \mathbb{E}((X - \mathbb{E}X) \times (Y - \mathbb{E}Y)) \\
 &= \mathbb{E}(XY - X\mathbb{E}(Y) - \mathbb{E}(X)Y + \mathbb{E}(X)\mathbb{E}(Y)) \\
 &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y) \\
 &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)
 \end{aligned}$$

■

Définition 115 (variables corrélées).

X et Y sont **corrélées** si $\text{cov}(X, Y) \neq 0$.

X et Y sont **corrélées positivement** si $\text{cov}(X, Y) > 0$.

X et Y sont **corrélées négativement** si $\text{cov}(X, Y) < 0$.

Proposition 116. X et Y indépendantes **implique** que $\text{cov}(X, Y) = 0$.

DÉMONSTRATION.

X et Y indépendantes



théorème 97 page 35

$$\mathbb{E}(XY) = \mathbb{E}(X) \times \mathbb{E}(Y)$$



proposition 114 page précédente

$$\text{cov}(X, Y) = 0 \blacksquare$$

Exemple 117 (variables continues dépendantes mais de covariance nulle). Soit $X \sim \mathcal{U}([-1, 1])$. Soit $S = \begin{cases} 1 & \text{avec probabilité } 1/2 \\ -1 & \text{avec probabilité } 1/2 \end{cases}$ qui indépendantes de X . On pose $Y = S \times X$.

1. X et Y sont dépendantes. On a $\mathbb{P}(X \in [-1/2, 1/2]) = \mathbb{P}(Y \in [-1/2, 1/2]) = 1/2$. Du coup on a :

$$\mathbb{P}(X \in [-1/2, 1/2]) = \mathbb{P}(X \in [-1/2, 1/2] \text{ et } Y \in [-1/2, 1/2]) \neq \mathbb{P}(X \in [-1/2, 1/2]) \times \mathbb{P}(Y \in [-1/2, 1/2])$$

(1/2 \neq 1/4)

2. Calculons la covariance de X et Y :

$$\begin{aligned} \text{cov}(X, Y) &= \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y \\ &= \mathbb{E}(XY) - 0\mathbb{E}Y \\ &= \mathbb{E}(XY) \\ &= \mathbb{E}(XSX) \\ &= \mathbb{E}(S)\mathbb{E}(X^2) \text{ car } S \text{ et } X^2 \text{ sont indépendantes (lemme des coalitions 43)} \\ &= 0 \times \mathbb{E}(X^2) = 0 \end{aligned}$$

Exemple 118 (variables discrètes dépendantes mais de covariance nulle). Soit $X \sim \mathcal{U}(\{-1, 0, 1\})$. Soit $Y = \begin{cases} 1 & \text{si } X = 0 \\ 0 & \text{sinon} \end{cases}$

1. X et Y sont dépendantes. En effet, on a $\mathbb{P}(X = 0) = 1/3$ et $\mathbb{P}(Y = 0) = 1/3$. Donc, $\mathbb{P}(X = 0) \times \mathbb{P}(Y = 0) = 1/9$. Mais $\mathbb{P}(X = 0 \text{ et } Y = 0) = 0$.

2. Calculons la covariance de X et Y :

$$\begin{aligned} \text{cov}(X, Y) &= \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y \\ &= 0 - 0\mathbb{E}Y \\ &= 0 \end{aligned}$$

5.4 Variance d'une somme

Proposition 119.

$$\begin{aligned} \mathbb{V}(X + Y) &= \mathbb{V}(X) + \mathbb{V}(Y) + 2\text{cov}(X, Y). \\ \mathbb{V}(\sum_{i=1}^n X_i) &= \sum_{i=1}^n \mathbb{V}(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{cov}(X_i, X_j). \end{aligned}$$

DÉMONSTRATION.

$$\begin{aligned} \mathbb{V}(X + Y) &= \mathbb{E}((X + Y)^2) - (\mathbb{E}(X + Y))^2 \\ &= \mathbb{E}(X^2 + Y^2 + 2XY) - (\mathbb{E}X + \mathbb{E}Y)^2 \\ &= \mathbb{E}(X^2) + \mathbb{E}(Y^2) + 2\mathbb{E}(XY) - [(\mathbb{E}X)^2 + (\mathbb{E}Y)^2 + 2\mathbb{E}X\mathbb{E}Y] \\ &= \underbrace{\mathbb{E}(X^2) - (\mathbb{E}X)^2}_{\mathbb{V}(X)} + \underbrace{\mathbb{E}(Y^2) - (\mathbb{E}Y)^2}_{\mathbb{V}(Y)} + 2 \underbrace{(\mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y)}_{\text{cov}(X, Y)} \\ &= \mathbb{V}(X) + \mathbb{V}(Y) + 2\text{cov}(X, Y) \end{aligned}$$

■

Corollaire 120.

Si X et Y sont non corrélées, alors $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y)$.

Si les X_1, \dots, X_n sont non corrélées deux à deux, alors $\mathbb{V}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathbb{V}(X_i)$.

5.5 Inégalités

5.5.1 Inégalité de Markov

Théorème 121 (inégalité de Markov). Soit X avec $\mathbb{P}(X \geq 0) = 1$. Pour tout $t > 0$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}$$

DÉMONSTRATION.

$$\mathbb{E}(X) = \int_{\mathbb{R}} x d\mathbb{P}_X(x) \geq \int_{[t, +\infty[} x d\mathbb{P}_X(x) \geq \int_{[t, +\infty[} t d\mathbb{P}_X(x) = t\mathbb{P}(X \geq t).$$

■

5.5.2 Inégalité de Bienaymé-Tchebychev

Théorème 122 (inégalité de Bienaymé-Tchebychev). Soit X de variance finie (et donc d'espérance finie). Pour tout $a > 0$,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathbb{V}(X)}{a^2}$$

DÉMONSTRATION.

$$\begin{aligned} \mathbb{P}(|X - \mathbb{E}(X)| \geq a) &= \mathbb{P}(|X - \mathbb{E}(X)|^2 \geq a^2) \\ &\leq \frac{\mathbb{E}(|X - \mathbb{E}(X)|^2)}{a^2} && \text{inégalité de Markov avec la va } |X - \mathbb{E}(X)|^2 \\ &= \frac{\mathbb{V}(X)}{a^2} \end{aligned}$$

■

5.5.3 Inégalité de Jensen

Théorème 123. Soit f une fonction convexe d'un intervalle réel I dans \mathbb{R} , et X une variable aléatoire à valeurs dans X . Si $\mathbb{E}(f(X))$ existe alors

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X)).$$

Aller plus loin

Voir les exemples de corrélation ici : https://www.lemonde.fr/les-decodeurs/article/2019/03/01/correlations-ou-c-5430063_4355770.html

Exercices

Dessiner les fonctions de répartition des lois usuelles

Calculer la probabilité d'avoir autant de pile que de face sur $2n$ lancers d'une pièce équilibrée

Calculer l'espérance du nombre de lancers d'une pièce équilibrée en moyenne pour avoir autant de pile que de face

Calculer les espérances et variances des différentes lois.

Montrer que la somme de deux lois binomiales est toujours binomiale. Explication intuitive ?

Montrer que la somme de deux variables suivant lois de Poisson est toujours de Poisson. Explication intuitive ?

Montrer que la somme de deux variables de lois normales, si indépendantes, suivent aussi une loi normale.

Montrer que si $T \sim \mathcal{E}(1)$, alors $\lceil \nu T \rceil \sim \mathcal{G}(1 - e^{1/\nu})$.

Donner un exemple de X et Y de covariance nulle mais qui sont dépendantes.

Si X et Y sont indépendantes, est-ce que X^2 et Y le sont aussi ?

Si X^2 et Y sont indépendantes, est-ce que X et Y le sont aussi ?

A-t-on $\mathbb{E}(g(X)) = g(\mathbb{E}(X))$?

Montrer que la discrétisation d'une loi exponentielle est une loi géométrique

Exercice 5.1. *Au casino, un joueur décide de continuer à miser jusqu'à ce qu'il gagne. Sa mise initiale est $a > 0$. A chaque partie, il double sa mise. A chaque partie, il a une probabilité p de gagner, et cela lui rapporte k fois la mise, $k \in \mathbb{N}^*$. Calculer l'espérance du gain G du joueur.*

Exercice 5.2. *Une urne contient une boule blanche et une boule noire.*

1. *On effectue des tirages avec remise jusqu'à obtention d'une boule blanche. Déterminer la loi de probabilité du nombre de N de tirages, puis calculer $\mathbb{E}(N)$ et $\mathbb{V}(N)$.*
2. *Mêmes questions si on remet une boule noire en plus après chaque tirage d'une boule noire. Calculer alors $\mathbb{P}(N > n)$ pour tout $n \in \mathbb{N}^*$.*

Chapitre 6

Simulation de lois

Dans ce chapitre, nous allons étudier comment simuler des lois sur un ordinateur. Voulez-vous une fonction qui tire un nombre selon une loi exponentielle ? Vous voulez-vous une fonction qui tire de manière uniforme un point dans un disque ? Allons-y ! C'est parti.

6.1 Loi uniforme sur $[0, 1]$

- Mersenne Twister
- Blum Blum Shub

6.2 Méthode d'inversion

Dans cette section, nous allons regarder comment concevoir une fonction qui tire aléatoirement un nombre selon une loi dont on connaît la fonction de répartition. C'est ce que l'on appelle la **méthode d'inversion**.

entrée : une fonction de répartition F

sortie : un simulateur d'une variable X de loi donnée par la fonction de répartition F

Dans cette méthode fonctionne quand F est inversible et que nous avons une expression de l'inverse de F . Elle fonctionne encore en utilisant l'inverse continue à gauche si jamais F n'est pas inversible. Par souci pédagogique, commençons par regarder le cas où F est inversible.

6.2.1 Cas où F est inversible

Si l'on dispose d'une expression, ou tout au moins d'algorithme pour calculer $F^{-1}(u)$ à partir de $u \in]0, 1[$, alors on peut tirer aléatoirement un nombre selon la loi de fonction de répartition F .

// retourne un nombre tiré aléatoirement selon la loi de fonction de répartition F

```
fonction simulerX()  
|   tirer  $u$  uniformément dans  $]0, 1[$   
|   renvoie  $F^{-1}(u)$ 
```

Proposition 124. *L'algorithme de simulation simulerX est correct.*

DÉMONSTRATION.

Soit X la variable retournée par l'algorithme simulerX . Autrement dit $X = F^{-1}(U)$ avec $U \sim \mathcal{U}_{[0,1]}$. Montrons que $F_X = F$.

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \leq x) \\ &= \mathbb{P}(F^{-1}(U) \leq x) \\ &= \mathbb{P}(U \leq F(x)) \\ &= F_U(F(x)) \\ &= F(x) \end{aligned}$$

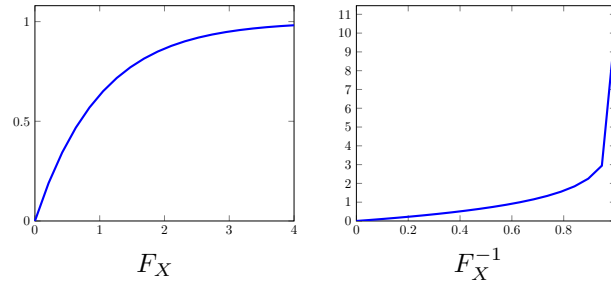


Exemple 125. On considère une variable X de loi exponentielle de paramètre λ :

$$X \sim \mathcal{E}(\lambda).$$

$$\text{On a } F_X(x) = 1 - e^{-\lambda x}.$$

$$F_X^{-1}(u) = -\frac{1}{\lambda} \ln(1 - u).$$



// retourne un nombre tiré aléatoirement selon la loi $\mathcal{E}(\lambda)$

```

fonction simulerLoiExponentielle( $\lambda$ )
|   tirer  $u$  uniformément dans  $[0, 1]$ 
|   renvoie  $-\frac{\ln(1-u)}{\lambda}$ 

```

Remarquons que si on tire u uniformément dans $[0, 1]$, alors $1 - u$ suit aussi une loi uniforme sur $[0, 1]$. On peut donc utiliser l'algorithme suivant :

// retourne un nombre tiré aléatoirement selon la loi $\mathcal{E}(\lambda)$

```

fonction simulerLoiExponentielle( $\lambda$ )
|   tirer  $u$  uniformément dans  $[0, 1]$ 
|   renvoie  $-\frac{\ln u}{\lambda}$ 

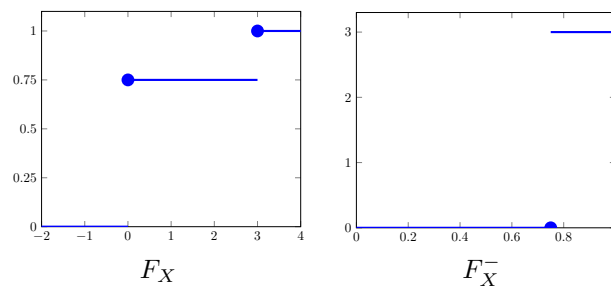
```

6.2.2 Cas général

Au lieu de prendre l'inverse F^{-1} qui n'existe pas forcément, on considère la **fonction inverse généralisée** F^- de F . Il s'agit de l'inverse continue à gauche de F : pour tout $u \in]0, 1[$,

$$F^-(u) = \inf \{x \in \mathbb{R} \mid F(x) \geq u\}.$$

Exemple 126. Voici la fonction de répartition d'une variable X avec $\mathbb{P}(X = 0) = 0.75$ et $\mathbb{P}(X = 3) = 0.25$:



Remarque 127. Si F bijective, alors $F^- = F^{-1}$.

Proposition 128. $F^-(u) \leq x$ ssi $u \leq F(x)$.

DÉMONSTRATION.

\Rightarrow Soit $x_0 = F^-(u)$. On a $F(x_0) \geq u$ car F est continue à gauche. On a $x_0 \leq x$. Comme F est croissante, $F(x) \geq F(x_0) \geq u$.

\Leftarrow

$u \leq F(x)$. Donc l'infimum $x_0 = F^-(u)$ est encore plus petit que x . ■

Proposition 129. $F(F^-(u)) = u$ si F est continue en $F^-(u)$.

```

fonction simulerX()
|   tirer  $u$  uniformément dans  $[0, 1]$ 
|   renvoie  $F^-(u)$ 

```

Théorème 130. *L'algorithme est correct.*

DÉMONSTRATION.

Soit $X = F^-(U)$ avec $U \sim \mathcal{U}_{[0,1]}$. On montre que $F_X = F$.

$$\begin{aligned}
 F_X(x) &= \mathbb{P}(X \leq x) \\
 &= \mathbb{P}(F^-(U) \leq x) \\
 &= \mathbb{P}(U \leq F(x)) \\
 &= F_U(F(x)) \\
 &= F(x)
 \end{aligned}$$

■

Vu l'algorithme précédent, on peut reformuler cela en l'existence d'une loi pour une fonction qui vérifie les propriétés d'une fonction de répartition.

Corollaire 131. *Soit F une fonction de \mathbb{R} dans $[0, 1]$, croissante, continue à droite avec $F(x) \xrightarrow{x \rightarrow -\infty} -1$ et $F(x) \xrightarrow{x \rightarrow +\infty} 1$. Alors il existe une mesure de probabilité sur \mathbb{R} dont la fonction de répartition est F .*

6.3 Loi normale

☹ Malheureusement, on ne peut pas utiliser la méthode d'inversion car **nous n'avons pas d'expression de l'inverse de la fonction de répartition pour une loi normale.**

Voici une solution pour générer la loi normale. Il s'agit de l'algorithme de Box-Muller. Ce dernier provient d'une interprétation probabiliste du calcul de l'intégrale de Gauss (voir proposition 70). Cet algorithme est en particulier utilisé dans `std::normal_distribution` de la librairie Standard C++.

On considère deux variables X et Y de loi normale centrée réduite indépendantes. On a :

$$\begin{aligned}
 p_X(x) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\
 p_Y(y) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}
 \end{aligned}$$

Comme X et Y sont indépendantes, on peut considérer le couple (X, Y) qui admet comme densité le produit des densités :

$$\begin{aligned}
 p_{X,Y}(x, y) &= p_X(x) \times p_Y(y) \\
 &= \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}
 \end{aligned}$$

L'idée est maintenant d'utiliser les coordonnées polaires, comme on l'avait fait dans la démonstration dans la proposition 70. Comme (X, Y) est un point aléatoire, on peut parler de ses coordonnées polaires (R, Θ) qui sont elle-même des variables aléatoires. On a :


$$\begin{aligned}
 X &= R \cos \Theta \\
 Y &= R \sin \Theta
 \end{aligned}$$

Étant donné la densité de (X, Y) , on voit que l'angle Θ est uniforme :

$$\Theta \sim \mathcal{U}(0, 2\pi).$$

Pour le rayon, c'est plus subtil. Étudions la distribution de R . Pour cela, on calcule la fonction de répartition de R (calcul similaire à ce qui était fait dans la démonstration de la proposition 70) :

$$\begin{aligned}
 F_R(r) &= \mathbb{P}(R \leq r) \\
 &= \int_{\text{disque de rayon } r} p_{X,Y}(x,y) dx dy \\
 &= \int_0^r \int_0^{2\pi} \frac{1}{2\pi} e^{-r^2/2} r dr d\theta \\
 &= \int_0^r r e^{-r^2/2} dr d\theta \\
 &= \int_0^r r^2/2 e^{-t} dt \\
 &= 1 - e^{-r^2/2}
 \end{aligned}$$

 **Proposition 132** (Algorithme de Box-Muller). Soit R une variable aléatoire de fonction répartition


$$F_R(r) = 1 - e^{-r^2/2}$$

et $\Theta \sim \mathcal{U}(0, 2\pi)$ avec R et Θ indépendantes. Alors

$$\begin{aligned}
 R \cos(\Theta) &\sim \mathcal{N}(0, 1) \\
 R \sin(\Theta) &\sim \mathcal{N}(0, 1)
 \end{aligned}$$

et sont indépendantes.

De manière équivalente, l'algorithme de Box-Muller est souvent reformulé comme cel a :

 **Proposition 133** (Algorithme de Box-Muller). Soit $T \sim \mathcal{E}(1)$ et $\Theta \sim \mathcal{U}(0, 2\pi)$ avec R et Θ indépendantes. Alors

$$\begin{aligned}
 \sqrt{2T} \cos(\Theta) &\sim \mathcal{N}(0, 1) \\
 \sqrt{2T} \sin(\Theta) &\sim \mathcal{N}(0, 1)
 \end{aligned}$$

et sont indépendantes.

DÉMONSTRATION.

On pose $T = R^2/2$. Autrement dit $R = \sqrt{2T}$. On a alors :

$$\begin{aligned}
 F_T(t) &= \mathbb{P}(T \leq t) \\
 &= \mathbb{P}(R^2/2 \leq t) \\
 &= \mathbb{P}(R \leq \sqrt{2t}) \\
 &= 1 - e^{-\sqrt{2t}^2/2} \\
 &= 1 - e^{-t}
 \end{aligned}$$

On reconnaît la fonction de répartition de la loi exponentielle $\mathcal{E}(1)$ de paramètre 1.

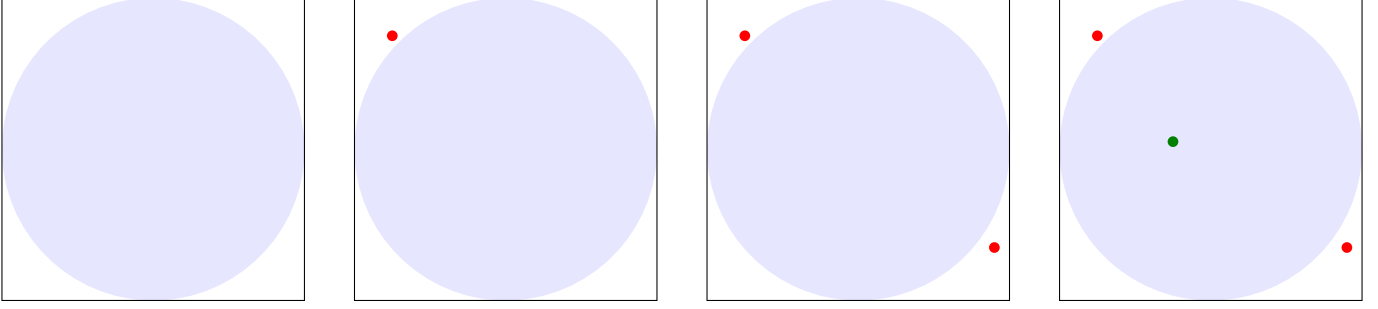
■

6.4 Méthode du rejet

6.4.1 Cas simple : loi uniforme sur le disque unité

On expose ici la méthode du rejet en montrant comment générer un point selon la loi uniforme sur le disque unité.

fonction générer un point dans le disque unité de manière uniforme avec la méthode du rejet
 | générer de manière uniforme un point x dans $[-1, 1]^2$
jusqu'à ce que x soit dans le disque unité
renvoie x



Proposition 134. *Le nombre d'itérations suit une loi géométrique de paramètre $\frac{\pi}{4}$. Son espérance est donc $\frac{4}{\pi}$.*

DÉMONSTRATION.

Un succès arrive quand le nombre généré est dans le disque unité. La probabilité d'avoir un succès est de $\text{aire}(\text{disque}) = \frac{\pi}{4}$.

À chaque tour de boucle, soit il a succès et on arrête, soit on échoue et on recommence. L'algorithme continue à itérer quand on pioche un point en dehors du disque unité. Bref, il s'agit du temps jusqu'au succès. ■

Proposition 135. *L'algorithme est correct.*

DÉMONSTRATION.

Démonstration simple mais un peu fausse et non rigoureuse. Soit X la variable aléatoire correspondant à l'algorithme. Etant donné un borélien A inclus dans le disque unité, on vérifie que $\mathbb{P}(X \in A)$ est égale à la mesure de A rapporté au disque unité.

$$\begin{aligned} \mathbb{P}(\text{valeur renvoyée} \in A) &= \mathbb{P}(X \in A \mid X \in \text{disque unité}) \\ &= \frac{X \in A \text{ et } X \in \text{disque unité}}{X \in \text{disque unité}} \\ &= \frac{\text{mesure de } A}{\text{surface du disque unité}} \end{aligned}$$

Version sérieuse. On note D le disque unité. Soit X_1, X_2, \dots les nombres générés par l'algorithme. On a $X_1, X_2, \dots \sim \mathcal{U}([-1, 1]^2)$ et iid. Soit T le premier instant de succès, i.e.

$$T = \min(t = 1, 2, \dots \mid X_t \in D)$$

La valeur renvoyée par l'algorithme est X_T . Calculons :

$$\begin{aligned} \mathbb{P}(X_T \in A) &= \mathbb{P}(X_1 \in A \text{ ou } (X_1 \notin D \text{ et } X_2 \in A) \text{ ou } (X_1 \notin D \text{ et } X_2 \notin D \text{ et } X_3 \in A) \text{ ou } \dots) \\ &= \sum_{i=1}^{\infty} \mathbb{P}(X_i \in A \text{ et } X_1, X_2, \dots, X_{i-1} \notin D) \\ &= \sum_{i=1}^{\infty} \mathbb{P}(X_i \in A) \times \mathbb{P}(X_1 \notin D) \times \dots \times \mathbb{P}(X_{i-1} \notin D) \\ &= \sum_{i=1}^{\infty} \text{mesure de } A \times \text{aire en dehors de } D^{i-1} \\ &= \text{mesure de } A \times \sum_{i=1}^{\infty} \text{aire en dehors de } D^{i-1} \\ &= \text{mesure de } A \times \sum_{i=0}^{\infty} \text{aire en dehors de } D^i \\ &= \text{mesure de } A \times \frac{1}{1 - \text{aire en dehors de } D} \\ &= \frac{\text{mesure de } A}{\text{surface du disque unité}} \end{aligned}$$

■

$$\begin{aligned}
\mathbb{P}(\text{valeur renvoyée} \leq x_0) &= \sum_{n \in \mathbb{N}^*} \mathbb{P}(X \leq x_0 \text{ et } Y \leq p(X)) \times \prod_{k=1}^{n-1} \mathbb{P}(\text{on n'a pas gardé le point } (X_k, Y_k)) \\
&= \sum_{n \in \mathbb{N}^*} \mathbb{P}(X \leq x_0 \text{ et } Y \leq p(X)) \times \mathbb{P}(\text{on n'a pas gardé le point } (X, Y))^{n-1} \\
&= \mathbb{P}(X \leq x_0 \text{ et } Y \leq p(X)) \times \frac{1}{1 - \mathbb{P}(\text{on n'a pas gardé le point } (X, Y))} \\
&= \frac{\mathbb{P}(X \leq x_0 \text{ et } Y \leq p(X))}{\mathbb{P}(Y \leq p(X))}.
\end{aligned}$$

Primo, $\mathbb{P}(X \leq x_0 \text{ et } Y \leq p(X)) = \frac{\text{surface sous la courbe jusqu'à } x_0}{\text{surface du rectangle } [a, b] \times [0, M]} = \frac{\int_{-\infty}^{x_0} p(x) dx}{M(b-a)}$.

Deuxio, $\mathbb{P}(Y \leq p(X)) = \frac{\text{surface sous la courbe}}{\text{surface du rectangle } [a, b] \times [0, M]} = \frac{1}{M(b-a)}$.

■

Théorème 137. *Le nombre d'itérations de l'algorithme avant d'obtenir le résultat suit une loi géométrique $\mathcal{G}(\frac{1}{M})$ où M est la borne.*

DÉMONSTRATION.

■



Cette méthode requiert que p soit à support compact et bornée.

6.4.4 Algorithme du rejet généralisé

Arrêtons maintenant de supposer que p n'est plus à support compact. On suppose maintenant que l'on dispose d'une autre fonction de densité g dont on sait générer des tirages de la loi correspondante et tel qu'il existe un réel $a \in \mathbb{R}_+$, pour tout $x \in \mathbb{R}$, $p(x) \leq ag(x)$.

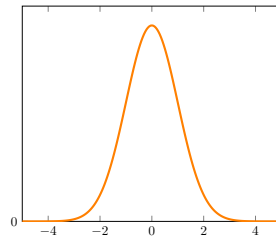
Remarquons que $a > 1$ car p et g sont des fonctions de densité (d'intégrale égale à 1). Le cas $a = 1$ signifie que $p = g$ mais ce cas là est stupide!

entrée :

- une fonction de densité p ;
- une fonction de densité g telle que l'on peut générer des nombres selon la loi de densité g ;
- un nombre $a > 1$ avec $p(x) \leq ag(x)$.

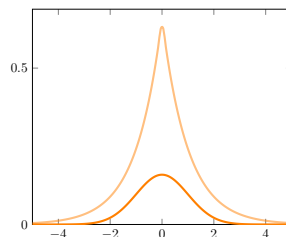
sortie : un générateur pour la distribution de densité p

Exemple 138. *Supposons que l'on veuille générer des tirages d'une loi normale centrée réduite. Soit $X \sim \mathcal{N}(0, 1)$. Voici la fonction de densité $p_{\mathcal{N}(0,1)}$ correspondante :*



On montre que $p_{\mathcal{N}(0,1)}(x) \leq \frac{e^{1/2-|x|}}{\sqrt{2\pi}}$ car $\frac{x^2}{2} - |x| + 1/2 = \frac{(|x|-1)^2}{2} \geq 0$.

Ainsi, $p_{\mathcal{N}(0,1)}(x) \leq \sqrt{\frac{2e}{\pi}} \frac{1}{2} e^{-|x|}$. On reconnaît $g(x) = \frac{1}{2} e^{-|x|}$ qui est la densité d'une loi double exponentielle (aussi appelée loi de Laplace) dont on sait faire des tirages. En orange clair, la fonction de densité de la loi de Laplace.



Pour trouver la valeur de a , on prend la fonction φ constante égale à 1 partout, et on obtient :

$$\mathbb{E}(1(X_n)1_{E_n}) = \mathbb{E}(1_{E_n}) = \frac{1}{a} \int_{-\infty}^{+\infty} p(x)dx = \frac{1}{a} = \alpha$$

Ainsi :

$$\begin{aligned} \mathbb{E}(\varphi(\hat{X})) &= \alpha \int_{-\infty}^{+\infty} \varphi(x)p(x)dx \sum_{n=1}^{\infty} (1-\alpha)^{n-1} \\ &= \int_{-\infty}^{+\infty} \varphi(x)p(x)dx \end{aligned}$$

■

Exercices

Exercice 6.2. On se propose d'étudier deux méthodes pour générer des points uniformément sur la sphère unité de dimension d .

1. Comment adapter l'algorithme de rejet donnée pour le disque unité, pour la sphère unité ?
2. Soit $Y_1, \dots, Y_d \sim \mathcal{N}(0, 1)$ iid. Quel est la loi de $Y/\|Y\|_2$?

Exercice 6.3. On souhaite simuler la loi $\mu = p_1\delta_1 + \dots + p_k\delta_k$ sur $\{1, \dots, k\}$.

1. Soit $U \sim \mathcal{U}([0, 1])$. Montrer que $X = \mathbf{1}_{U < p_1} + 2\mathbf{1}_{p_1 \leq U < p_1 + p_2} + \dots + k\mathbf{1}_{p_1 + \dots + p_{k-1} \leq U < 1}$ soit la loi μ .
2. Expliquer le lien avec la méthode d'inversion.
3. Écrire un algorithme qui simule la loi $\mu = p_1\delta_1 + \dots + p_k\delta_k$ sur $\{1, \dots, k\}$.

Exercice 6.4. 1. Si X suit une loi exponentielle de paramètre λ , quelle est la loi de $\lfloor X \rfloor$?

2. En déduire un algorithme de simulation de la loi géométrique de paramètre $p \in]0, 1[$ (cf. définition 54 page 25), grâce à une variable aléatoire uniforme sur $[0, 1]$.

Exercice 6.5. Soit $(T_k)_{k \in \mathbb{N}^*}$ iid de loi $\mathcal{E}(\lambda)$. On note $S_n = T_1 + \dots + T_n$. On définit $N = \sum_{k=1}^{\infty} \mathbf{1}_{S_k \leq 1}$.

1. Quelle est l'interprétation de T_n ?
2. Quelle est l'interprétation de N ?
3. Montrer que S_n suit la loi Gamma de densité $x \mapsto \lambda e^{-\lambda x} \frac{(\lambda x)^{n-1}}{(n-1)!} \mathbf{1}_{x > 0}$.
4. Montrer que les événements $S_n \leq 1$ et $N \geq n$ sont égaux.
5. Calculer $\mathbb{P}(N = n)$.
6. En déduire que N suit une loi de Poisson de paramètre λ .

Chapitre 7

Convergence de suites de v.a.

7.1 Notation pour la moyenne

Voici la notation classique pour la moyenne. D'une suite $(X_n)_{n \in \mathbb{N}}$, on définit la suite des moyennes, i.e. $(\overline{X}_n)_{n \in \mathbb{N}}$ où \overline{X}_n est la moyenne des n premiers termes X_1, \dots, X_n .

Notation 140 (moyenne). La moyenne de X_1, \dots, X_n est $\frac{1}{n} \sum_{k=1}^n X_k$ et se note \overline{X}_n .

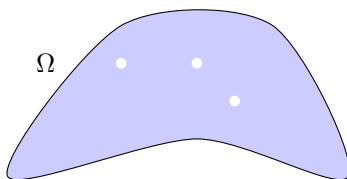
Remarque 141. Attention, la barre est juste sur le X . On note \overline{X}_n et non pas $\overline{X_n}$. En effet, c'est comme si on avait une suite $X = (X_n)_{n \in \mathbb{N}}$, et on définit la suite des moyennes, i.e. $\overline{X} = (\overline{X}_n)_{n \in \mathbb{N}}$. La notation $\overline{X_n}$ serait impropre car la moyenne $\overline{X_n}$ ne dépend pas juste de X_n !

7.2 Différents types de convergence

7.2.1 Convergence presque sûre

Définition 142 (convergence presque sûre). $X_n \xrightarrow[n \rightarrow +\infty]{p.s.} X$ si $\mathbb{P}(X_n \xrightarrow[n \rightarrow \infty]{} X) = 1$.

Dit plus précisément, $X_n \xrightarrow[n \rightarrow +\infty]{p.s.} X$ lorsque $\mathbb{P}(\{\omega \in \Omega \mid X_n(\omega) \xrightarrow[n \rightarrow \infty]{} X(\omega)\}) = 1$, c'est-à-dire quand les mondes possibles ω où $X_n(\omega)$ ne converge pas vers $X(\omega)$ forment un événement de mesure nulle. Voici un dessin de l'espace Ω . Sur toute la zone bleu de probabilité 1, X_n converge vers X quand n tend vers $+\infty$. Juste en quelques points (en blanc), la suite $(X_n)_{n \in \mathbb{N}}$ fait autre chose.



Voici un exemple conséquence de la loi forte des grands nombres (théorème 162 page 54).

Exemple 143. Soit $X_1, \dots, X_n, \dots \sim \mathbb{B}(p)$ iid. Alors

$$\overline{X}_n \xrightarrow[n \rightarrow +\infty]{p.s.} p.$$

7.2.2 Convergence en probabilité

Définition 144 (convergence en probabilité). $X_n \xrightarrow[n \rightarrow +\infty]{\mathbb{P}} X$ si pour tout $\epsilon > 0$, $\mathbb{P}(|X_n - X| \geq \epsilon) \xrightarrow[n \rightarrow \infty]{} 0$.

Loi du zéro-un de Borel

On présente ici l'artillerie pour montrer que la suite donnée en exemple ci-dessous, qui converge en proba, ne converge pas presque sûrement. Etant donné une suite d'événements A_0, A_1, A_2, \dots , la définition suivante définit l'événement qui dit qu'il y a une infinité d'indices n tel que A_n se soit produit.

Définition 145. Soit $(A_n)_{n \geq 0}$ une suite d'événements. La **limite supérieure** des A_n est

$$\limsup_{n \rightarrow \infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k.$$

Le théorème suivant, relie la convergence de la série $\sum_{n \in \mathbb{N}} \mathbb{P}(A_n)$ et la probabilité de $\mathbb{P}(\limsup_n A_n)$ qui vaut soit 0 soit 1.

Théorème 146 (loi du zéro-un de Borel). Soit $(A_n)_{n \geq 0}$ une suite d'événements **indépendants**.

1. Si $\sum_{n \in \mathbb{N}} \mathbb{P}(A_n)$ converge, alors $\mathbb{P}(\limsup_n A_n) = 0$;
2. Sinon, si $\sum_{n \in \mathbb{N}} \mathbb{P}(A_n)$ diverge, alors $\mathbb{P}(\limsup_n A_n) = 1$.

Le point 1 de la loi du zéro-un fonctionne aussi si les événements ne sont pas indépendants. Il s'agit du lemme de Borel-Cantelli.

Lemme 147 (de Borel-Cantelli). Soit $(A_n)_{n \geq 0}$ une suite d'événements. Si $\sum_{n \in \mathbb{N}} \mathbb{P}(A_n)$ est convergente, alors $\mathbb{P}(\limsup_n A_n) = 0$.

Exemple 148 (convergence en proba mais pas presque sûre). Soit $(X_n)_{n \in \mathbb{N}}$ toutes indépendantes avec $\mathbb{P}(X_n = n) = \frac{1}{n}$ et $\mathbb{P}(X_n = 0) = 1 - \frac{1}{n}$, i.e. X_n est une sorte de Bernouilli de paramètre $\frac{1}{n}$ mais où la valeur de succès est n .

- On a $X_n \xrightarrow[n \rightarrow +\infty]{\mathbb{P}} 0$. En effet, $\mathbb{P}(|X_n| \geq \epsilon) = \frac{1}{n}$ pour n suffisamment grand, et tend vers 0 quand n tend vers $+\infty$.
- Par contre, X_n ne converge pas presque sûrement.
Si convergence presque sûre il doit y avoir, c'est vers 0 (se montre par l'absurde).
Ici, $\sum_{n \in \mathbb{N}} \mathbb{P}(X_n = n) = \sum_{n \in \mathbb{N}} \frac{1}{n}$ est divergente. D'après la loi du zéro-un de Borel, $\mathbb{P}(\limsup_n (X_n = n)) = 1$.
L'événement $\limsup_n (X_n = n)$ signifie qu'il y a une infinité d'indices n avec $X_n = n$. Ainsi, cela implique que la suite $(X_n(\omega))_{n \in \mathbb{N}}$ prend des valeurs arbitraires grandes avec une probabilité de 1. Bref, presque sûrement, X_n ne converge pas vers 0 : on a $\mathbb{P}(X_n \xrightarrow[n \rightarrow \infty]{} 0) = 0$.

7.2.3 Convergence en loi

On cherche à modéliser le fait que la loi de X_n ressemble de plus en plus à la loi de X quand n tend vers l'infini.

Définition 149 (convergence en loi). $X_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} X$ si pour tout fonction continue bornée f , on a

$$\mathbb{E}(f(X_n)) \xrightarrow[n \rightarrow \infty]{} \mathbb{E}(f(X)).$$

Proposition 150 (convergence en loi pour des variables à valeurs entières). Soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables à valeurs dans \mathbb{N} et X à valeur dans \mathbb{N} . On a :

$$X_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} X \text{ ssi pour tout } k \in \mathbb{N}, \mathbb{P}(X_n = k) \xrightarrow[n \rightarrow \infty]{} \mathbb{P}(X = k).$$

Théorème porte-manteau

On suppose que les variables aléatoires sont à valeurs dans E , et le théorème porte-manteau donne des caractérisations de la convergence en loi.

Définition 151 (adhérence). *L'adhérence de A , notée \overline{A} , est le plus petit fermé qui inclut A .*

Définition 152 (intérieur). *L'intérieur de A , notée $\overset{\circ}{A}$, est le plus grand ouvert inclus dans A .*

Définition 153 (frontière de A). *La frontière de A , notée ∂A est $\overline{A} \setminus \overset{\circ}{A}$.*

Théorème 154. *Sont équivalentes :*

1. $X_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} X$
2. pour toute fonction uniformément continue bornée f , $\mathbb{E}(f(X_n)) \xrightarrow[n \rightarrow \infty]{} \mathbb{E}(f(X))$
3. Pour fermé F de E , $\limsup_n \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F)$
4. Pour ouvert O de E , $\liminf_n \mathbb{P}(X_n \in O) \geq \mathbb{P}(X \in O)$
5. pour tout borélien A de E tel que $\mathbb{P}(X \in \partial A) = 0$, $\mathbb{P}(X_n \in A) \xrightarrow[n \rightarrow \infty]{} \mathbb{P}(X \in A)$

Théorème 155. *Soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires et X à valeurs réelles. Soit F_{X_n} la fonction de répartition de X_n et F_X la fonction de répartition F_X . Alors sont équivalentes :*

- $X_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} X$;
- pour tout point de continuité x de F_X , on a $F_{X_n}(x) \xrightarrow[n \rightarrow \infty]{} F_X(x)$.

Exemple 156. *Une suite de variables aléatoires X_n qui suit une loi normale centrée en 0 et de variance $\frac{1}{n}$ converge en loi vers X de loi le Dirac en 0 :*

$$\mathcal{N}(0, \frac{1}{n}) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \delta_0.$$

Il y a convergence point à point de F_{X_n} vers F_X , autrement dit, pour tout $x \neq 0$, $F_{X_n}(x) \xrightarrow[n \rightarrow \infty]{} F_X(x)$. Par contre, il n'y a pas de convergence au point de discontinuité en 0 de F_X . En effet, $F_{X_n}(0) = \frac{1}{2}$ alors que $F_X(0) = 1$.

Proposition 157. *Soit f une fonction continue. Si $X_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} X$, alors $f(X_n) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} f(X)$.*

7.2.4 Comparaison des types de convergence

Théorème 158 (implication des convergences).

$$\begin{aligned} X_n \xrightarrow[n \rightarrow +\infty]{p.s.} X &\implies X_n \xrightarrow[n \rightarrow +\infty]{\mathbb{P}} X \implies X_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} X \\ X_n \xrightarrow[n \rightarrow +\infty]{\mathbb{P}} a &\iff X_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} a \text{ (Dirac en } a) \end{aligned}$$

7.3 Théorème de Slutsky


Théorème 159 (Théorème de Slutsky).

$$X_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} X \text{ et } Y_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} c \quad \text{implique} \quad \begin{cases} X_n + Y_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} X + c \\ X_n \times Y_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} cX \end{cases}$$

Exercice 7.1. *A-t-on $X_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} X$ et $Y_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} Y$ implique $X_n + Y_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} X + Y$?
Et et si les X_n et Y_n sont indépendants ?*

7.4 Loi des grands nombres

Voici une version faible de la loi des grands nombres, faible car les hypothèses sont justement trop fortes. Ce théorème est pédagogiquement intéressant car sa démonstration est une belle application de l'inégalité de Bienaymé–Chebyshev (théorème 122 page 39).

 **Théorème 160** (loi faible des grands nombres). *Si les $(X_n)_n$, telles X_1, X_2, \dots ont toutes la même espérance μ et une variance finie σ^2 . De plus on suppose qu'elles soient deux à deux décorrélées. Alors*

$$\overline{X}_n \xrightarrow[n \rightarrow +\infty]{\mathbb{P}} \mu$$

DÉMONSTRATION.

Pour tout $n \in \mathbb{N}$, on a par linéarité de l'espérance

$$\mathbb{E}(\overline{X}_n) = \frac{1}{n} \mathbb{E}(X_1 + \dots + X_n) = \frac{1}{n} (\mathbb{E}(X_1) + \dots + \mathbb{E}(X_n)) = \frac{n}{n} \mathbb{E}(X_0) = \mu \quad (7.1)$$

Comme les X_1, X_2, \dots sont deux à deux non corrélées, comme vu dans le corollaire 120 on a

$$\mathbb{V}(X_1 + X_2 + \dots + X_n) = \sum_{i=1}^n \mathbb{V}(X_i)$$

Ainsi,

$$\begin{aligned} \mathbb{V}(\overline{X}_n) &= \mathbb{V}\left(\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right) \\ &= \frac{1}{n^2} \mathbb{V}(X_1 + X_2 + \dots + X_n) \\ &= \frac{1}{n^2} \sum_{i=1}^n \mathbb{V}(X_i) \\ &= \frac{n}{n^2} \sigma^2 \\ &= \sigma^2/n \end{aligned}$$

Soit $\epsilon > 0$. On a :

$$\begin{aligned} \mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) &= \mathbb{P}(|\overline{X}_n - \mathbb{E}(\overline{X}_n)| \geq \epsilon) \\ &\leq \frac{\mathbb{V}(\overline{X}_n)}{\epsilon^2} \text{ via l'inégalité de Bienaymé-Tchebychev (cf. théorème 122)} \\ &\leq \frac{1}{n} \times \sigma^2 / \epsilon \\ &\xrightarrow[n \rightarrow \infty]{} 0 \end{aligned}$$

■

Remarque 161. *On peut imaginer des versions plus subtiles de la loi faible des grands nombres. Par exemple, les espérances de chaque X_n peuvent être différentes, mais converge vers un certain μ . Pareil, on a supposé que les variances sont toutes égales à σ^2 , mais en fait on a juste besoin que $\frac{1}{n^2} \sum_{i=1}^n \mathbb{V}(X_i)$ tend vers 0 quand n tend vers $+\infty$. Ces versions ne sont pas intéressantes en pratique.*

Ce qui nous importe plus, c'est le type de convergence. La version forte de la loi des grands nombres est une convergence presque sûre. Voici un résultat plus général, démontré par Etemadi. C'est ce résultat là que l'on retiendra. On n'a pas d'hypothèse sur la . Sa démonstration est plus compliquée.

Théorème 162 (loi forte des grands nombres). *Si les $(X_n)_{n \in \mathbb{N}}$ sont deux à deux indépendants et de même loi qui admet un moment d'ordre 1 alors*

$$\overline{X}_n \xrightarrow[n \rightarrow +\infty]{p.s.} \mathbb{E}(X_0).$$

7.5 Théorème central limite

Voici **un des plus beaux théorèmes de l'histoire des mathématiques**. Oui, les lois normales résument les moyennes de variables iid. Sa démonstration, une fois, que l'on a bien posé les bases des transformées de Fourier (fonction caractéristique) n'est pas difficile. Mais c'est technique, donc elle est étudiée dans le chapitre 19.

Théorème 163 (théorème central limite). Soit $(X_n)_n$ iid admettant un moment d'ordre 2. On note μ l'espérance et σ^2 la variance de X_n . Alors :

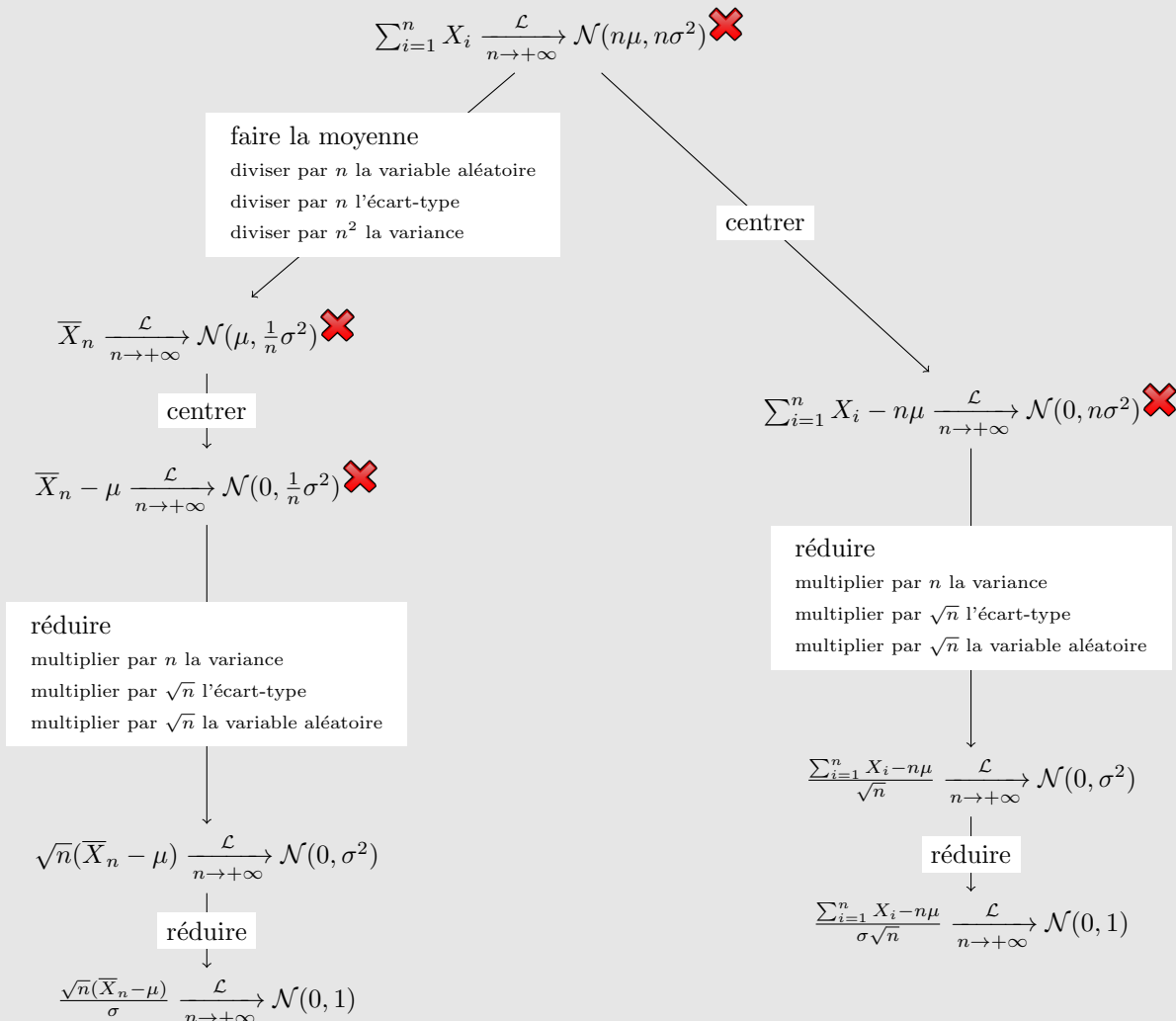
$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2).$$

On verra le théorème central limite dans cas vectoriel, voir Théorème 206.

Différentes formulations du théorème centrale limite

Voici plusieurs formulations du théorème centrale limite, avec aussi des formulations mathématiquement incorrectes mais pédagogiquement intéressantes (marquées par **✗**).

Par linéarité de l'espérance, la moyenne de $\sum_{i=1}^n X_i$ est $n\mu$. La variance de $\sum_{i=1}^n X_i$ est $n\sigma^2$ (comme les variables sont indépendantes, on somme les variances comme montré dans la corollaire 120). La première formulation, mathématiquement fausse car n apparaît en partie droite, est de dire que la limite en loi de $\sum_{i=1}^n X_i$ est $\mathcal{N}(n\mu, n\sigma^2)$. Partant de là, en faisant la moyenne, en centrant, en réduisant, on obtient d'autres formulations du théorème central limite.



L'exemple qui suit montre l'utilisation du TCL pour calculer une probabilité, alors qu'une méthode directe a un grand coût de calcul!

Exemple 164.



On considère une variable aléatoire S_n qui suit une loi binomiale $\mathcal{B}(n, \frac{1}{2})$. Calculons $\mathbb{P}(S_{2000} > 1111)$. Une première méthode consiste à appliquer la définition de la loi binomiale :

$$\mathbb{P}(S_{2000} > 1111) = \sum_{k=1112}^{2000} \binom{2000}{k} \left(\frac{1}{2}\right)^{2000}$$

Le calcul est pénible...

Une seconde méthode est que S_n suit la même loi qu'une somme $\sum_{i=1}^n X_n$ de n variables aléatoires X_1, \dots, X_n de Bernoulli iid d'espérance $p = \frac{1}{2}$. On rappelle que la variance σ^2 d'une loi de Bernoulli d'espérance $\frac{1}{2}$ est $p(1-p) = \frac{1}{4}$. D'après le TCL, on a :

$$\frac{\sum_{i=1}^n X_n - \frac{n}{2}}{\frac{\sqrt{n}}{2}} \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

Ainsi, $\frac{S_{2000} - \frac{2000}{2}}{\frac{\sqrt{2000}}{2}}$ suit quasiment la loi $\mathcal{N}(0, 1)$.

$$\begin{aligned} \mathbb{P}(S_{2000} > 1111) &= \mathbb{P}(S_{2000} - 1000 > 111) \\ &= \mathbb{P}\left(\frac{S_{2000} - 1000}{\sqrt{2000}/2} > \frac{111}{\sqrt{2000}/2}\right) \\ &= \mathbb{P}\left(Y > \frac{111}{\sqrt{500}}\right) \end{aligned}$$

$\mathbb{P}\left(Y > \frac{111}{\sqrt{500}}\right)$ s'approxime à l'aide du TCL par la probabilité qu'une variable aléatoire de loi normale réduite centrée soit $> \frac{111}{\sqrt{500}}$. Ainsi, on donne une expression à l'aide de la densité de la loi normale réduite centrée :

$$\mathbb{P}\left(Y > \frac{111}{\sqrt{500}}\right) \approx \int_{\frac{111}{\sqrt{500}}}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx.$$

7.6 D'autres convergences en loi

On relate ici deux convergences en loi que l'on a déjà évoqué dans la section 4.4.2 et section 4.3.5. Ayant maintenant défini proprement le concept de convergence en loi, on peut revisiter les rapprochement entre lois binomiales et loi de Poisson d'une part, et entre les lois géométriques et les lois exponentielles d'autre part.

7.6.1 Loi des événements rares

!“Loi” dans “Loi des événements rares” signifie que la loi de Poisson modélise le **nombre de succès** pour les événements rares, dans le sens donné par le théorème suivant.

Théorème 165. Soit $(p_n)_{n \in \mathbb{N}}$ une suite avec $np_n \xrightarrow[n \rightarrow +\infty]{} \lambda \in \mathbb{R}$. On a :

$$\mathcal{B}(n, p_n) \xrightarrow[x \rightarrow +\infty]{\mathcal{L}} \mathcal{P}(\lambda).$$

7.6.2 Comparaison loi géométrique et loi exponentielle

On s'intéresse ici au **temps d'attente** avant d'obtenir un succès.

Théorème 166. Soit $(p_n)_{n \in \mathbb{N}}$ et $(a_n)_{n \in \mathbb{N}}$ avec $p_n \xrightarrow[n \rightarrow \infty]{} 0$ et $\frac{p_n}{a_n} \xrightarrow[n \rightarrow \infty]{} \lambda > 0$. Alors :

$$\text{Si } Y_n \sim \mathcal{G}(p_n) \text{ alors } a_n Y_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{E}(\lambda).$$

Exercices

Supposons que la durée pour avoir succès soit une loi exponentielle. Montrer que le nombre de succès entre l'instant 0 et T soit une loi de Poisson.

Pour les plus mathématiciens, étudier https://en.wikipedia.org/wiki/Vague_topology

Chapitre 8

Vecteurs aléatoires

Définition 167. Un vecteur aléatoire est une application $X : \Omega \rightarrow \mathbb{R}^k$. C'est un vecteur $\begin{pmatrix} X_1 \\ \dots \\ X_k \end{pmatrix}$ où les X_i sont des variables aléatoires réelles.

Exemple 168 (lancers de dé). On considère n lancers indépendants d'un même dé. On considère le vecteur

$$N = \begin{pmatrix} N_1 \\ N_2 \\ N_3 \\ N_4 \\ N_5 \\ N_6 \end{pmatrix} \text{ avec } \begin{array}{l} N_1 := \text{nombre de } \square \text{ obtenus sur les } n \text{ lancers} \\ N_2 := \text{nombre de } \square \text{ obtenus sur les } n \text{ lancers} \\ N_3 := \text{nombre de } \square \text{ obtenus sur les } n \text{ lancers} \\ N_4 := \text{nombre de } \square \text{ obtenus sur les } n \text{ lancers} \\ N_5 := \text{nombre de } \square \text{ obtenus sur les } n \text{ lancers} \\ N_6 := \text{nombre de } \square \text{ obtenus sur les } n \text{ lancers} \end{array}$$

Bien sûr, $N_1 + N_2 + \dots + N_6 = n$. Par exemple, pour un monde possible ω avec les 7 lancers $\square \square \square \square \square \square \square$, on a

$$X(\omega) = \begin{pmatrix} 3 \\ 1 \\ 0 \\ 0 \\ 2 \\ 1 \end{pmatrix}.$$

8.1 Espérance

L'espérance d'un vecteur est juste l'espérance coordonnée par coordonnée. C'est le vecteur où la i -ème coordonnée contient l'espérance de X_i , i.e. l'espérance de la i -coordonnée de X .

Définition 169. L'espérance du vecteur aléatoire $X = \begin{pmatrix} X_1 \\ \dots \\ X_k \end{pmatrix}$ est $\mathbb{E}(X) := \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_k) \end{pmatrix}$.

Exemple 170 (lancers de dé). $\mathbb{E}(N) = n \begin{pmatrix} 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \end{pmatrix}$.

8.2 Matrice de variance-covariances

8.2.1 Définition

La matrice $\mathbb{V}(X)$ de covariance de X est la matrice contenant la covariance de X_i et X_j à la ligne i et colonne j , et la variance de X_i sur la i -ème case de la diagonale.

$$\begin{pmatrix} \mathbb{V}(X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_{n-1}) & \mathbb{V}(X_n) \\ \text{cov}(X_1, X_2) & \mathbb{V}(X_2) & \cdots & \text{cov}(X_2, X_{n-1}) & \text{cov}(X_2, X_n) \\ \vdots & & & & \vdots \\ \text{cov}(X_1, X_n) & & & & \mathbb{V}(X_n) \end{pmatrix}$$

Définition 171 (matrice de covariance). $\mathbb{V}(X) = (\text{cov}(X_i, X_j))_{i,j=1..n}$.

Exemple 172 (lancers de dé). Chaque N_i soit une loi binomiale $\mathcal{B}(n, 1/6)$. La variance est de $n \frac{1}{6} (1 - \frac{1}{6}) = n \frac{5}{36}$.

On calcule la covariance $\text{Cov}(N_i, N_j)$ en étudiant $N_i + N_j$:

— D'une part, $N_i + N_j \sim \mathcal{B}(n, 2/6)$. En effet, avoir un succès c'est tombé sur i ou j . Donc $\mathbb{V}(N_i + N_j) = n \frac{2}{6} (1 - \frac{2}{6}) = n \frac{8}{36}$.

— D'autre part, $\mathbb{V}(N_i + N_j) = \mathbb{V}(N_i) + \mathbb{V}(N_j) + 2\text{Cov}(N_i, N_j) = n \frac{10}{36} + 2\text{Cov}(X_i, X_j)$.

Ainsi, $\text{Cov}(X_i, X_j) = -n \frac{1}{36}$. On obtient donc la matrice de covariances suivante :

$$\mathbb{V}(X) = n \frac{1}{36} \begin{pmatrix} 5 & -1 & -1 & -1 & -1 & -1 \\ -1 & 5 & -1 & -1 & -1 & -1 \\ -1 & -1 & 5 & -1 & -1 & -1 \\ -1 & -1 & -1 & 5 & -1 & -1 \\ -1 & -1 & -1 & -1 & 5 & -1 \\ -1 & -1 & -1 & -1 & -1 & 5 \end{pmatrix}.$$

$\text{Cov}(N_i, N_j)$ est négative car si N_i augmente, ça a tendance à faire baisser N_j . Sur un nombre fixe de lancers de dé, avoir plus de 1 a tendance à faire baisser le nombre de 2 par exemple.

8.2.2 Matrice symétrique

Définition 173 (matrice symétrique). Une matrice M est symétrique si $M_{ij} = M_{ji}$.

Proposition 174. Toute matrice de covariance est symétrique.

8.2.3 Matrice définie semi-positive

Matrice définie semi-positive

Pour une variable aléatoire réelle (cas 1D), la variance est positive (ou nulle). Pour un vecteur aléatoire (de dimension k quelconque), la variance est généralisée par la **matrice de covariance**. Ainsi le fait que la **variance soit positive ou nulle** est généralisée par le fait que cette matrice soit **définie semi-positive**.

Définition 175 (matrice définie semi-positive). Une matrice symétrique M est définie semi-positive si pour tout vecteur $x \in \mathbb{R}^k$ on a $x^\top \cdot M \cdot x \geq 0$.

On donne une écriture matricielle de la matrice de covariance afin de faciliter la démonstration qu'elle est définie semi-positive.

Proposition 176 (écriture matricielle). $\mathbb{V}(X) = \mathbb{E}(RR^\top)$ avec $R = X - \mathbb{E}(X)$.

DÉMONSTRATION.

A la ligne i , colonne j , nous avons $\mathbb{V}(X)_{ij} = \mathbb{E}((X_i - \mathbb{E}(X_i)) \times (X_j - \mathbb{E}(X_j))^t) = \text{cov}(X_i, X_j)$. Sur la diagonale, nous avons $\mathbb{V}(X)_{ii} = \mathbb{V}(X_i)$. ■

Théorème 177. Une matrice de variance-covariances est définie semi-positive.

DÉMONSTRATION.

$\mathbb{V}(X)$ s'écrit $\mathbb{E}(RR^\top)$ où $R = X - \mathbb{E}(X)$.

Ainsi

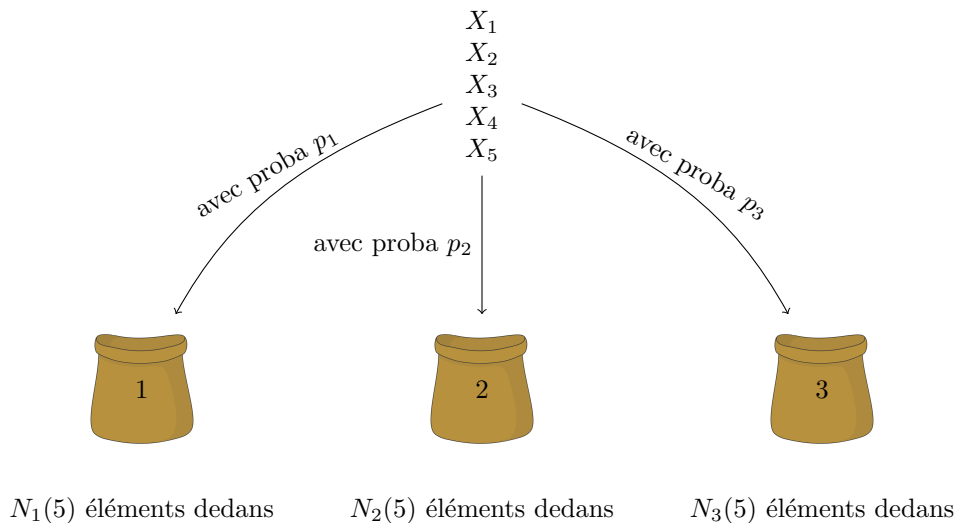
$$\begin{aligned} x^\top \cdot \mathbb{V}(X) \cdot x &= \mathbb{E}(x^\top R R^\top x) \\ &= \mathbb{E}((R^\top x)^2) \geq 0 \end{aligned}$$

■

8.3 Loi multinomiale

La loi binomiale compte le nombre d'éléments qui sont dans le paquet 1 si la probabilité de mettre un élément vaut p (loi de Bernoulli de paramètre p).

On considère maintenant des variables discrètes X_1, X_2, \dots, X_n indépendantes à valeurs dans $\{1, \dots, k\}$ (k résultats au dé) où la probabilité d'être dans la classe $i \in \{1, \dots, k\}$ vaut p_i , le vecteur \mathbf{p} est un **vecteur de probabilité**, i.e. $\sum_{i=1}^k p_i = 1$. Pour n éléments aléatoires, on compte alors le nombre $N_i(n)$ d'éléments qui vont être dans la classe i .



Chaque variable $N_i(n)$ soit la loi binomiale de paramètre n et p_i : cela correspond à la loi de Bernoulli de paramètre p_i avec les choix “aller dans le paquet i ” (succès) et “ne pas aller dans le paquet i ” (échec).

Proposition 178. $N_i(n) \sim \mathcal{B}(n, p_i)$.

La loi multinomiale, elle, concerne, le vecteur $N(n) := \begin{pmatrix} N_1(n) \\ \vdots \\ N_k(n) \end{pmatrix}$.

Définition 179. Soit $\mathbf{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_k \end{pmatrix}$ une loi de probabilité sur $\{1, \dots, k\}$. Soit $X_1, \dots, X_n \sim \mathbf{p}$ iid.

On pose $N_i(n) := \#\{j \in \{1, \dots, n\} \mid X_j = i\}$. Le vecteur $N(n) := \begin{pmatrix} N_1(n) \\ \vdots \\ N_k(n) \end{pmatrix}$ suit la **loi multinomiale** de paramètre (n, \mathbf{p}) .

Exemple 180 (lancers de dé). $X_1, \dots, X_n \sim \begin{pmatrix} 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \end{pmatrix}$.

$N_1(n) :=$ nombre de obtenus sur les n lancers

$N_2(n) :=$ nombre de obtenus sur les n lancers

$N_3(n) :=$ nombre de obtenus sur les n lancers

$N_4(n) :=$ nombre de obtenus sur les n lancers

$N_5(n) :=$ nombre de obtenus sur les n lancers

$N_6(n) :=$ nombre de obtenus sur les n lancers

Exemple 181 (vote). n étudiants votent pour trois candidats pour la présidence de leur syndicat. Soient $N_1(n), N_2(n), N_3(n)$ les nombres de votes correspondants, et supposons que les n étudiants votent indépendamment avec des probabilités $p_1 = 0.45$, $p_2 = 0.4$, et $p_3 = 0.15$. Trouver la loi conjointe de X_1, X_2, X_3 . Calculer la loi marginale de X_3 , et la loi conditionnelle de X_1 sachant $X_3 = 3$.

Dénombrement et loi multinomiale

$p_1^{n_1} \times \cdots \times p_k^{n_k}$ est la probabilité d'avoir un certain tirage avec n_1 éléments dans le paquet 1, n_2 éléments dans le paquet 2, etc. et n_k éléments dans le paquet k . Par exemple, $p_1^{n_1} \times \cdots \times p_k^{n_k}$ est la probabilité d'avoir les n_1 premiers éléments dans le paquet 1, les n_2 éléments suivants dans le paquet 2, etc. puis n_k dernier éléments dans le paquet k .

Par exemple pour $k = 3$, $n_1 = 5$, $n_2 = 2$, $n_3 = 4$:

11111223333

$p_1^{n_1} \times \cdots \times p_k^{n_k}$ est aussi la probabilité d'avoir

21311311233.

Combien y-a-t-il de façon de permuter ces éléments ?

— On choisit d'abord les n_1 emplacements pour les 1 parmi les $n = n_1 + n_2 + \dots n_k$ possibilités ;

?1??1?1?1?1

— On choisit ensuite les n_2 emplacements pour les 2 parmi les emplacements restants, à savoir les $n_2 + \dots n_k$ restants ;

21??121?1?1

— etc.

— Finalement, les n_k emplacements restants sont clairs : ils sont pour les n_k emplacements pour les k .

21331213131

Voici l'expression qui donne le nombre de façon de les ordonner :

$$\binom{n_1 + n_2 + \dots + n_k}{n_1} \binom{n_2 + \dots + n_k}{n_2} \cdots \binom{n_k}{n_k} = \frac{n!}{n_1! n_2! \dots n_k!}$$

Proposition 182.

$$\mathbb{P}(N(n) = \begin{pmatrix} n_1 \\ \vdots \\ n_k \end{pmatrix}) = \begin{cases} \frac{n!}{n_1! \times \dots \times n_k!} p_1^{n_1} \times \cdots \times p_k^{n_k} & \text{si } n_1 + \dots + n_k = n \\ 0 & \text{sinon.} \end{cases}$$

DÉMONSTRATION.

Cf. cadre gris plus haut ! ■

Proposition 183. $\frac{N(n)}{n} \xrightarrow[n \rightarrow +\infty]{p.s.} \mathbf{p}$.

DÉMONSTRATION.

On le démontre avec la loi des grands nombres. On a :

$$\frac{N_i(n)}{n} = \frac{1}{n} \sum_{t=1}^n Z_t$$

où $Z_t = \begin{cases} 1 & \text{si au temps } t, \text{ l'objet choisi va dans le paquet } i. \\ 0 & \text{sinon.} \end{cases}$

Comme les Z_t sont d'espérance p_i , la loi des grands nombres donne :

$$\frac{N(n)_i}{n} \xrightarrow[n \rightarrow +\infty]{p.s.} p_i.$$

Ainsi, $\mathbb{P}(\frac{N(n)_i}{n} \xrightarrow[n \rightarrow \infty]{} p_i) = 1$. Par le corollaire 9, on a :

$$\mathbb{P}(\bigcap_{i=1}^k \frac{N(n)_i}{n} \xrightarrow[n \rightarrow \infty]{} p_i) = 1$$

qui est le résultat escompté. ■

Proposition 184. $\mathbb{E}(N(n)) = np$.

DÉMONSTRATION.

$N_i(n)$ soit une loi binomiale $\mathcal{B}(n, p_i)$.

$$N_i(n) = \sum_{t=1}^n Z_t$$

où $Z_t = \begin{cases} 1 & \text{si au temps } t, \text{ l'objet choisi va dans le paquet } i. \\ 0 & \text{sinon.} \end{cases}$

Ainsi

$$\begin{aligned} \mathbb{E}(N_i(n)) &= \mathbb{E}\left(\sum_{t=1}^n Z_t\right) \\ &= \sum_{t=1}^n \mathbb{E}(Z_t) \\ &= \sum_{t=1}^n p_i \\ &= np_i. \end{aligned}$$

■

Proposition 185.

1. $\mathbb{V}(N(n))_{ii} = np_i(1 - p_i)$.
2. $\mathbb{V}(N(n))_{ij} = -np_i p_j$ si $i \neq j$.

DÉMONSTRATION.

1. Pour montrer que $\mathbb{V}(N(n))_{ii} = np_i(1 - p_i)$, on rappelle que

$$N_i(n) = \sum_{t=1}^n Z_t$$

où $Z_t = \begin{cases} 1 & \text{si au temps } t, \text{ l'objet choisi va dans le paquet } i. \\ 0 & \text{sinon.} \end{cases}$

Les variables Z_t sont mutuellement indépendantes. Elles sont donc deux à deux indépendantes. Elles sont deux à deux non corrélées. Donc via le corollaire 120 :

$$\begin{aligned} \mathbb{V}(N_i(n)) &= \sum_{t=1}^n \mathbb{V}(Z_t) \\ &= \sum_{t=1}^n p_i(1 - p_i) \\ &= np_i(1 - p_i) \end{aligned}$$

2. Pour montrer $\mathbb{V}(N(n))_{ij} = -np_i p_j$ si $i \neq j$, on rappelle la formule de la covariance :

$$\text{cov}(N_i(n), N_j(n)) = \frac{1}{2}[\mathbb{V}(X_i + X_j) - \mathbb{V}(X_i) - \mathbb{V}(X_j)]$$

Ainsi :

$$\text{cov}(N_i(n), N_j(n)) = \frac{1}{2}[n(p_i + p_j)(1 - p_i - p_j) - np_i(1 - p_i) - np_j(1 - p_j)] = -p_i p_j.$$

■

Ainsi, la matrice de variances-covariances ressemble à ça :

Corollaire 186.

$$\begin{aligned} \mathbb{V}(N(n)) &= n \times \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & p_1p_3 & \cdots & -p_1p_{k-1} & -p_1p_k \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 & \cdots & -p_2p_{k-1} & -p_2p_k \\ -p_3p_1 & -p_3p_2 & p_3(1-p_3) & \cdots & -p_3p_{k-1} & -p_3p_k \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -p_kp_1 & p_kp_2 & p_kp_3 & \cdots & -p_kp_{k-1} & p_k(1-p_k) \end{pmatrix} \\ &= n \times (\Delta_{\mathbf{p}} - \mathbf{p}\mathbf{p}^T) \end{aligned}$$

où $\Delta_{\mathbf{p}}$ est la matrice diagonale avec \mathbf{p} sur la diagonale.

8.4 Lois normales vectorielles

La loi normale réelle se généralise aux vecteurs, que l'on appelle vecteur gaussien.

8.4.1 Définition

Définition 187 (vecteur gaussien).

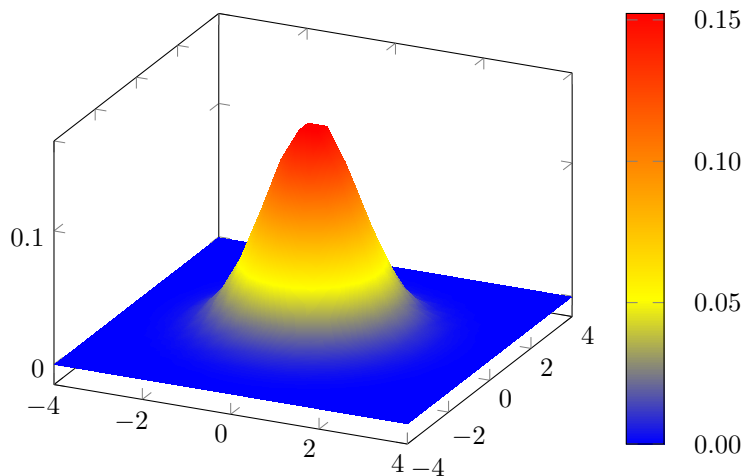
Un vecteur $X = (X_1, \dots, X_k)$ suit une loi normale vectorielle

si toute combinaison linéaire sur X_1, \dots, X_k suit une loi normale réelle, ou alors un Dirac.

Remarque 188. Le Dirac correspond à une loi normale réelle de variance nulle.

Exemple 189 (Densité d'un vecteur gaussien à deux dimensions).

L'image suivante est l'exemple typique d'une densité gaussienne :



Il s'agit d'un vecteur (X_1, X_2) où $X_1, X_2 \sim \mathcal{N}(0, 1)$, X_1 et X_2 sont indépendants. La fonction densité est :

$$p(x) = \frac{1}{(\sqrt{2\pi})^2} e^{-\frac{1}{2}(x_1^2 + x_2^2)}.$$

Exemple 190. Soit $X_1 \sim \mathcal{N}(0, 1)$ et $X_2 = X_1$. Alors le vecteur (X_1, X_2) est un vecteur gaussien.

Exemple 191. Soit $X_1 \sim \mathcal{N}(0, 1)$ et $X_2 = -X_1$. Alors le vecteur (X_1, X_2) est un vecteur gaussien.

Exemple 192 (contre-exemple). ! Il ne suffit pas que X_1, \dots, X_n suivent des lois normales pour que $X = (X_1, \dots, X_n)$ suive une loi normale vectorielle. Soit $X_1 \sim \mathcal{N}(0, 1)$.

$$\text{Soit } X_2 = \begin{cases} X_1 & \text{if } |X_1| \leq c \\ -X_1 & \text{if } |X_1| > c \end{cases}$$

Bien que X_1 et X_2 suivent tous les deux des lois normales, le vecteur (X_1, X_2) ne suit pas une loi normale vectorielle.

8.4.2 Caractérisation par espérance et matrice de covariance

Théorème 193. Pour tout $\mu \in \mathbb{R}^k$, pour toute matrice Σ matrice $k \times k$ définie semi-positive, on peut construire un vecteur gaussien X avec

$$\mathbb{E}(X) = \mu \text{ et } \mathbb{V}(X) = \Sigma.$$

Rappel d'algèbre linéaire : diagonalisation d'une matrice définie semi-positive

Toute matrice définie semi-positive Σ se diagonalise. Une matrice de covariance se diagonalise donc. Il existe une matrice O orthogonale et des réels $\lambda_1, \dots, \lambda_n \geq 0$ tels que :

$$\Sigma = U \text{diag}(\lambda_1, \dots, \lambda_n) U^\top.$$

Ainsi, Σ admet une sorte de **racine carrée** : une matrice A telle que $\Sigma = AA^\top$. Il s'agit de

$$A := U \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n}) U^\top.$$

C'est le pendant de l'**écart-type**.

Théorème 194. Deux vecteurs gaussiens ayant même espérance μ et même matrice de covariance Σ ont même loi.

Notation 195. On note

$$X \sim \mathcal{N}(\mu, \Sigma).$$

μ vecteur moyenne
 Σ matrice de variance-covariances

Théorème 196. Soit X un vecteur normal.

$$X_i \text{ et } X_j \text{ sont indépendants } \text{ssi } \text{Cov}(X_i, X_j) = 0.$$

Exemple 197. Un vecteur $X \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\right)$ est tel que

$$\begin{aligned} X_1 &\sim \mathcal{N}(0, 1) \\ X_2 &= 0 \text{ avec probabilité } 1. \end{aligned}$$

8.4.3 Densité quand la matrice de covariance est définie strictement positive

Matrice définie semi-positive et matrice définie strictement positive

Un nombre strictement positif est inversible, mais pas 0 : on ne peut pas diviser par 0. De même une matrice définie semi-positive n'est pas forcément inversible.

La notion de **nombre positif ou nul** se généralise par la notion de **matrice définie semi-positive**.
 La notion de **nombre strictement positif** se généralise par la notion de **matrice définie strictement positive**.

Définition 198 (matrice définie strictement positive). Une matrice symétrique M est définie positive si pour tout vecteur x non nul on a $x^\top \cdot M \cdot x > 0$.

Théorème 199. Soit $\mu \in \mathbb{R}^k$ et Σ matrice **définie strictement positive**. La densité de $\mathcal{N}(\mu, \Sigma)$ s'écrit :

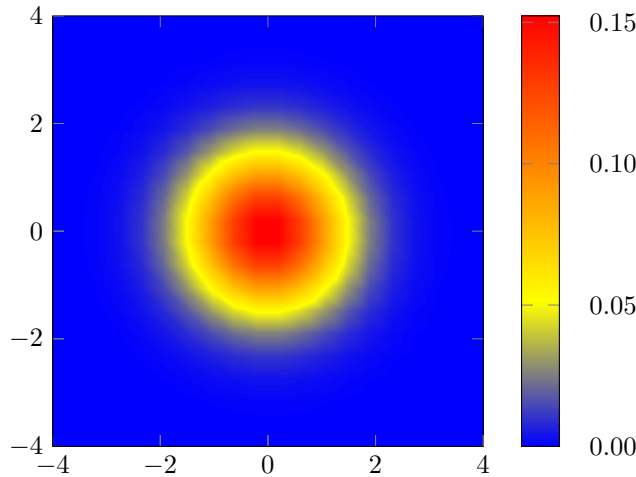
$$p_{\mathcal{N}(\mu, \Sigma)}(x) = \frac{1}{(\sqrt{2\pi})^k \sqrt{\det \Sigma}} e^{-\frac{1}{2}((x-\mu)^\top \Sigma^{-1} (x-\mu))}$$

Remarque 200. On notera la ressemblance avec la formule de la densité d'une loi normale réelle qui est

$$p_{\mathcal{N}(\mu, \sigma^2)}(x) = \frac{1}{\sqrt{2\pi} \sqrt{\sigma^2}} e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}}.$$

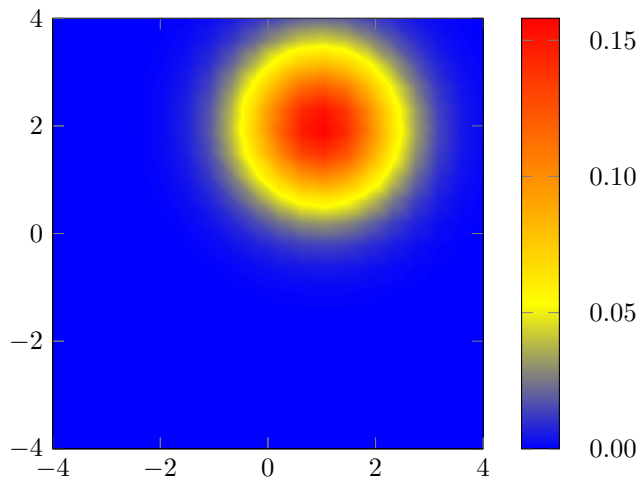
Les exemples qui suivent sont en 2D. On donne des heatmaps qui représentent la densité.

Exemple 201 (densité de $\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, Id_2\right)$). Voici la heatmap de la fonction de densité. De beaux cercles concentriques sont les lignes de niveau.



Exemple 202 (densité de $\mathcal{N}\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, Id_2\right)$).

Si l'espérance est $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$, alors la fonction de densité est juste translaté. Les cercles concentriques sont centrées en $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$.



Exemple 203 (densité de $\mathcal{N}\left((,) \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}\right)$). On considère $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ qui suit une loi $\mathcal{N}\left((,) \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}\right)$.

On pose $\Sigma = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$. On a deux vecteurs propres donnés par $\Sigma \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 4 \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ et $\Sigma \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

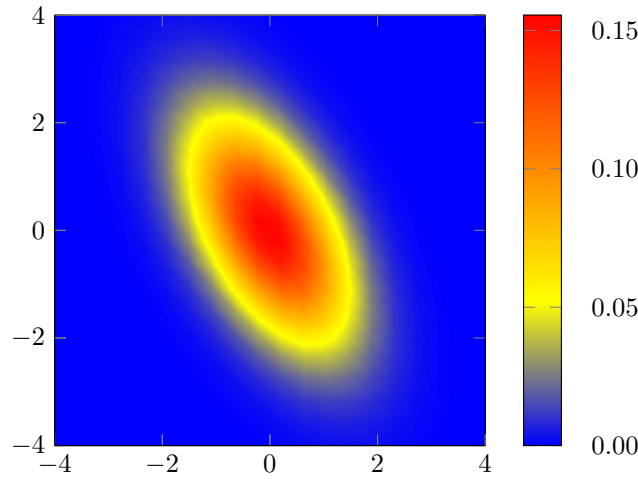
En posant $P = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, on a :

$$P^{-1}\Sigma P = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}.$$

Dans cet exemple, $X_1 + X_2$ et $X_1 - X_2$ sont décorrélées et indépendantes.

Exemple 204 (densité de $\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.75 \\ -0.75 & 2 \end{pmatrix}\right)$).

Considérons maintenant une matrice de covariance un peu plus chiadé. Considérons un vecteur $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.75 \\ -0.75 & 2 \end{pmatrix}\right)$. Voici la heatmap de la fonction de densité correspondante. Les variables X_1 et X_2 sont négativement corrélés car $\text{cov}(X_1, X_2) = -0.75$. C'est pourquoi nous avons des ellipses concentriques inclinées.



La diagonalisation de Σ permet de construire de trouver les axes des ellipses ; et donc des variables correspondantes décorréliées.

$$\Sigma = \begin{pmatrix} 1.0000 & -0.7500 \\ -0.7500 & 2.0000 \end{pmatrix}$$

$$\Sigma^{-1} = \begin{pmatrix} 1.3913 & 0.5217 \\ 0.5217 & 0.6957 \end{pmatrix}$$

$$U = \begin{pmatrix} -0.8817 & 0.4719 \\ -0.4719 & -0.8817 \end{pmatrix}$$

$$U^{-1} = \begin{pmatrix} -0.8817 & -0.4719 \\ 0.4719 & -0.8817 \end{pmatrix}$$

$$D = \begin{pmatrix} 0.5986 & 0.0000 \\ 0.0000 & 2.4014 \end{pmatrix}$$

$$D^{-1} = \begin{pmatrix} 1/0.5986 & 0.0000 \\ 0.0000 & 1/2.4014 \end{pmatrix}$$

$$U \times D \times U^{-1} = \begin{pmatrix} 1.0000 & -0.7500 \\ -0.7500 & 2.0000 \end{pmatrix} \quad \text{On retrouve } \Sigma$$

$$U \times D^{-1} \times U^{-1} = \begin{pmatrix} 1.3913 & 0.5217 \\ 0.5217 & 0.6957 \end{pmatrix} \quad \text{L'expression de } \Sigma \text{ diagonalisée permet de calculer } \Sigma^{-1} \text{ facilement : il suffit de calculer l'inverse de } D.$$

A priori, les variables $-0.8817X_1 + 0.4719X_2$ et $-0.4719X_1 - 0.8817X_2$ sont décorréliées.

Matrice de covariance. Elle contient l'information de combien s'éparpille les échantillons autour de la moyenne, ainsi que les directions principales. Bref, elle synthétise l'information des ellipses concentriques de la heatmap.

Inverse de la matrice de covariance. Son expression apparaît dans l'exposant de l'exponentielle de la fonction de densité.

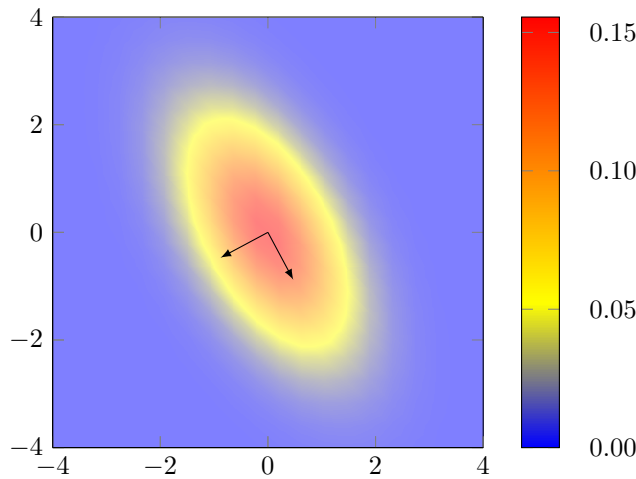
Matrice de passage : chaque colonne contient les coordonnées dans le repère traditionnel des vecteurs propres, i.e. les axes des ellipses concentriques dessinées dans la figure ci-dessous

Inverse de la matrice de passage : elle contient les coordonnées des vecteurs unités des axes (repère traditionnel) dans le repère des vecteurs propres

Matrice diagonale qui correspond à la matrice covariance mais dans la base des vecteurs propres. Sur le premier vecteur propre, la variance est de 0.5986, et sur le deuxième vecteur propre elle est de 2.4014.

Une matrice diagonale est facilement à inverser : on inverse les coefficients sur la diagonale.

Vecteurs propres de Σ représentés sur la heatmap :



8.4.4 Changement affine

Proposition 205. Soit $X \sim \mathcal{N}(\mu, \Sigma)$. Alors :

$$AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\top).$$

8.5 Théorème centrale limite vectoriel


Théorème 206 (théorème centrale limite vectoriel). Soit $(X_n)_n$ des vecteurs aléatoires de \mathbb{R}^k iid admettant un moment d'ordre 2. On note μ l'espérance et Σ la variance de X_n . Alors :

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma)$$

Une autre formulation du TCL vectoriel est :

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}} \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma).$$

On peut appliquer le TCL sur des vecteurs qui suivent une probabilité \mathbf{p} .

 **Corollaire 207.** Soit $N(n)$ de loi multinomiale de paramètre (n, \mathbf{p}) .

$$\sqrt{n}\left(\frac{N(n)}{n} - \mathbf{p}\right) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, \Delta_{\mathbf{p}} - \mathbf{p}\mathbf{p}^\top)$$

où $\Delta_{\mathbf{p}}$ est la matrice diagonale de diagonale \mathbf{p} .

DÉMONSTRATION.

Soit X_n qui vaut la variable aléatoire qui suit la loi donnée par le vecteur \mathbf{p} au temps n . On suppose que les X_n sont indépendantes, comme dans la définition d'une loi multinomiale, on a $N(n)_i = \sum_{j=1}^k 1_{X_n=i}$.

Considérons $(Y(n))_{n \in \mathbb{N}}$ où $Y_i(n) = 1_{X_n=i}$, et appliquons le TCL vectoriel. On obtient :

$$\sqrt{n}(\bar{Y}_n - \mathbb{E}(Y(1))) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, \mathbb{V}(Y(1)))$$

1. On a $\bar{Y}_n = \frac{N(n)}{n}$. En effet :

$$\begin{aligned}
\bar{Y}_n &= \frac{1}{n} \sum_{j=1}^n Y(j) \\
&= \frac{1}{n} \left(\sum_{j=1}^k Y_i(j) \right)_{i=1..k} \\
&= \frac{1}{n} \left(\sum_{j=1}^k 1_{X_n=i} \right)_{i=1..k} \\
&= \frac{1}{n} N(n)
\end{aligned}$$

2. Montrons que $\mathbb{E}(Y(1)) = \mathbf{p}$. On a :

$$\begin{aligned}
\mathbb{E}(Y(1)) &= (\mathbb{E}(Y(1)_i))_{i=1..k} \\
&= (\mathbb{P}(X_1 = i))_{i=1..k} \\
&= \mathbf{p}
\end{aligned}$$

3. Un peu avec les même calculs que proposition 185, on a $\mathbb{V}(Y(1)) = \Delta_{\mathbf{p}} - \mathbf{p}\mathbf{p}^T$.

■

Exemple 208 (nombre de piles et de faces). Le vecteur de probabilité est $\mathbf{p} = (1/2, 1/2)$.

La matrice $\Delta_{\mathbf{p}} - \mathbf{p}\mathbf{p}^T$ est

$$\Sigma = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix} - \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} (1/2, 1/2) = \begin{pmatrix} 1/4 & -1/4 \\ -1/4 & 1/4 \end{pmatrix} = 1/4 \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

On a

$$\Sigma = O \text{diag}(1/2, 0) O^{-1}$$

avec

$$O = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

La présence de la valeur propre nulle, fait que Σ n'est pas définie positive. Cela vient du fait que $N_1(n) + N_2(n) = n$.

Exercices

Soit X_1, \dots, X_k . Montrer que la matrice de covariance de X_1, \dots, X_k est inversible ssi la seule combinaison linéaire certaine de X_1, \dots, X_k est la combinaison nulle.

Démontrer que

$$\mathbb{P}(N(n) = \begin{pmatrix} n_1 \\ \vdots \\ n_k \end{pmatrix}) = \begin{cases} \frac{n!}{n_1! \times \dots \times n_k!} p_1^{n_1} \times \dots \times p_k^{n_k} & \text{si } n_1 + \dots + n_k = n \\ 0 & \text{sinon.} \end{cases}$$

en calculant directement $\mathbb{P}(N(n) = \begin{pmatrix} n_1 \\ \vdots \\ n_k \end{pmatrix})$ comme une intersection d'événements et en faisant apparaître des probabilités conditionnelles.

Deuxième partie

Statistiques

Chapitre 9

Introduction

9.1 Modèles statistiques

À partir de maintenant, la mesure de probabilité n'est plus connue. Nous allons faire des estimations pour en avoir des informations. On se place dans le cas paramétré. Autrement dit, on a un peu d'informations mais on ne connaît pas les paramètres.

Exemple 209. *On sait que la pièce de monnaie soit une Bernoulli $\mathcal{B}(p)$ mais on ne connaît pas le paramètre p .*

Ainsi, contrairement à un espace probabilisé, maintenant nous avons une **famille de mesure de probabilités**, paramétrée par θ . Le **paramètre** θ est inconnu. Il peut être un ou plusieurs nombres par exemple.

Exemple 210. *On sait que le phénomène suit une loi normale mais on ne connaît pas le paramètre $\theta = (\mu, \sigma^2)$. Autrement dit, la moyenne μ et la variance σ^2 sont inconnues.*

Définition 211 (modèle statistique). *Un modèle statistique est $(\Omega, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ est la donnée :*

- d'un espace Ω ;
- d'une tribu \mathcal{F} sur Ω ;
- d'une famille $(\mathbb{P}_\theta)_{\theta \in \Theta}$ de mesures de probabilité.

Définition 212. θ s'appelle le paramètre. Il est inconnu.

Dans la suite, on travaille avec l'espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$ paramétrée par le paramètre $\theta \in \Theta$. Dans la littérature, on note $\mathbb{P}_\theta(A)$ ou alors $\mathbb{P}(A; \theta)$ la probabilité de l'événement A quand le paramètre θ . De même, on note $\mathbb{E}_\theta(X)$ ou $\mathbb{E}(X; \theta)$ l'espérance de X quand le paramètre vaut θ . Parfois, on omet le paramètre θ , et on écrit simplement $\mathbb{P}(A)$ ou $\mathbb{E}(X)$.

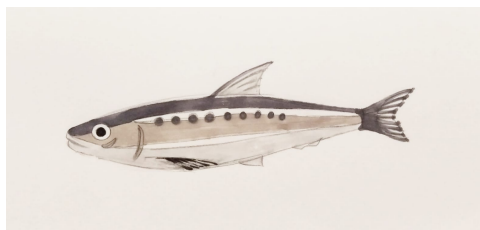
Définition 213 (statistique). *C'est juste une question de vocabulaire. Mais dans le contexte des statistiques, une variable aléatoire $X : \Omega \rightarrow \mathbb{R}$ s'appelle une **statistique**.*

9.2 Échantillon aléatoire

Définition 214 (échantillon aléatoire). *Un échantillon aléatoire est une suite (X_1, \dots, X_n) où pour tout $\theta \in \Theta$, X_1, \dots, X_n sont des variables aléatoires iid. de même loi sur \mathbb{P}_θ .*

Exemple 215 (pièce de monnaie où la probabilité d'avoir pile est inconnue). *Un pièce de monnaie où la probabilité θ de tomber sur pile est le paramètre **inconnu**. On considère un modèle statistique où, pour tout θ , les variables X_1, \dots, X_n sont iid et suivent une loi de Bernoulli de paramètre θ .*

Exemple 216 (longueur des sardines).



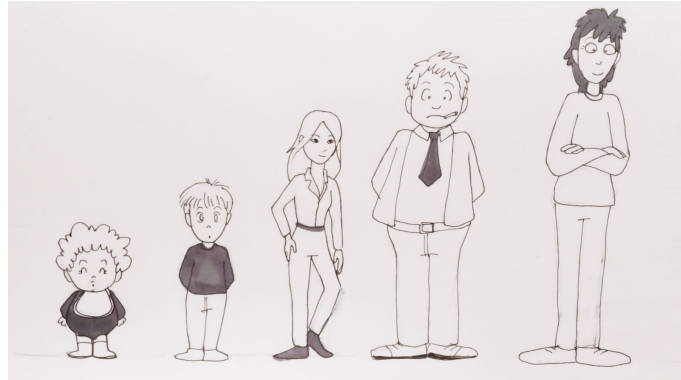
On suppose que la longueur de sardines est modélisée par une loi normale dont la moyenne μ et l'écart-type σ sont inconnus. Le paramètre est

$$\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+.$$

Ainsi, la taille X_i de la i -ème sardine vérifie $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

On considère donc X_1, \dots, X_n iid qui suivent chacune $\mathcal{N}(\mu, \sigma^2)$ où X_i est la taille de la i -ème sardine

Exemple 217 (hauteur de personnes).



On suppose que la hauteur des personnes dépend du sexe. Ainsi, on modélise selon par une **loi de mélange** de deux lois normales :

$$0.5\mathcal{N}(\mu_1, \sigma_1^2) + 0.5\mathcal{N}(\mu_2, \sigma_2^2)$$

Le paramètre est $\theta := (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$.

Chapitre 10

Estimation

Dans ce chapitre, on se fixe X_1, \dots, X_n un échantillon aléatoire, elles sont iid de même loi sur \mathbb{P}_θ .

Estimation

entrée : une suite de données x_1, \dots, x_n ;

sortie : une estimation du paramètre inconnu $\theta \in \Theta$.

Les données x_1, \dots, x_n sont typiquement des réalisations de X_1, \dots, X_n .

10.1 Exemples

Le paramètre peut être un nombre, mais aussi un vecteur de nombres :

- l'espérance
- la variance (l'espérance est par exemple connu et vaut par exemple 0)
- l'espérance et la variance

Pour du clustering (chapitre 11 page 79) :

- les centres des gaussiennes dans une mixture gaussienne (c'est le problème que résout l'algorithme des k -moyennes, les matrices de covariances sont supposées connues et valent Id)
- les centres des gaussiennes et les matrices de covariances (c'est le cadre de l'algorithme EM)

Pour la régression linéaire (chapitre 16 page 115) :

- les coefficients de la droite, du plan, ou de l'hyper-plan qui s'approche le plus des données

10.2 Estimateur

Un **estimateur** prend les données et estime le paramètre θ . Un estimateur peut être bon ou mauvais, là n'est pas encore la question. On verra ça dans la section suivante.

Exemple 218. On estime la moyenne par $\frac{x_1 + \dots + x_n}{n}$. Ainsi, on pourrait dire que l'estimateur est la fonction

$$(x_1, \dots, x_n) \mapsto \frac{x_1 + \dots + x_n}{n}.$$

Définition 219 (estimateur). Quand n est fixé, un **estimateur** de θ est une fonction $\mathbb{R}^n \rightarrow \mathbb{R}$.

Le nombre n de données est souvent variable. On sait pas combien on va faire de mesures. Ainsi, un estimateur est une collection de fonctions, une pour chaque nombre n de mesures.

Définition 220 (estimateur). Dans le contexte où n est variable, un **estimateur** de θ est une suite $(\hat{\theta}_n)_{n \in \mathbb{N}}$ de fonctions $\hat{\theta}_n : \mathbb{R}^n \rightarrow \mathbb{R}$.

Exemple 221. La suite $(\hat{\theta}_n)_{n \in \mathbb{N}}$ de fonctions $\hat{\theta}_n : (x_1, \dots, x_n) \mapsto \frac{x_1 + \dots + x_n}{n}$ est un estimateur de l'espérance θ .

Remarque 222. Par simplicité, on dira que $\hat{\theta}_n$ est un estimateur au lieu de $(\hat{\theta}_n)_{n \in \mathbb{N}}$. Aussi, quand le contexte est clair, on écrit $\hat{\theta}_n$ au lieu de $\hat{\theta}_n(X_1, \dots, X_n)$ où X_1, \dots, X_n est l'échantillon aléatoire.

Il arrive que nous voulions estimer une partie de θ . Par exemple, si $\theta = (\mu, \sigma^2)$ où μ est la moyenne de la loi et σ^2 la variance, on pourrait ne vouloir estimer que μ . C'est-à-dire que à la fois μ et σ^2 sont inconnus, mais on ne s'intéresse qu'à estimer μ . On pourrait aussi vouloir estimer autre chose, comme μ^2 . C'est pourquoi on peut estimer $g(\theta)$ où g est une fonction connue.

Exemple 223. Pour estimer la moyenne μ , on estime $g(\theta)$ où $g : (\mu, \sigma^2) \mapsto \mu$.

Pour éviter des surcharges de notation, on continue le reste de l'exposé en estimant θ .

10.3 Biais

Maintenant, on évalue la qualité d'un estimateur.

10.3.1 Estimateur sans biais

Définition 224 (estimateur sans biais). L'estimateur $\hat{\theta}_n$ est **sans biais** si pour tout $\theta \in \Theta$, $\mathbb{E}(\hat{\theta}_n(X_1, \dots, X_n)) = \theta$.

Dans la définition précédente, le “pour tout $\theta \in \Theta$ ” signifie que l'on fixe la valeur de θ pour tout θ . Une fois fixée, elle nous donne une mesure de probabilité, et donc un moyen de calculer $\mathbb{E}(\hat{\theta}_n)$. Un estimateur sans biais signifie que si le paramètre vaut θ , alors l'espérance de l'estimateur est θ (θ étant inconnu).

Exemple 225 (moyenne). Soit X_1, \dots, X_n un échantillon aléatoire dont la moyenne inconnue est θ . L'estimateur de la moyenne vérifie :

$$\mathbb{E}(\hat{\theta}_n(X_1, \dots, X_n)) = \mathbb{E}\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n}(\mathbb{E}(X_1) + \dots + \mathbb{E}(X_n)) = \frac{1}{n}(\theta + \dots + \theta) = \theta.$$

Ainsi, l'estimateur

$$(x_1, \dots, x_n) \mapsto \frac{x_1 + \dots + x_n}{n}$$

est sans biais.

10.3.2 Estimateur asymptotiquement sans biais

On se place ici dans le contexte où n est variable. Un estimateur asymptotiquement sans biais si l'estimateur tend à être sans biais plus on fait de mesures.

Définition 226 (estimateur asymptotiquement sans biais). L'estimateur $\hat{\theta}_n$ est **asymptotiquement sans biais** si pour tout $\theta \in \Theta$, $\mathbb{E}(\hat{\theta}_n) \xrightarrow{n \rightarrow +\infty} \theta$.

La définition précédente est un peu similaire à un estimateur sans biais, mais avec une convergence vers θ quand n tend vers l'infini. Cela veut dire que plus le nombre de données est grand, moins l'estimateur a de biais.


10.3.3 Estimateurs de la variance

Voici une première définition **naïve** d'un estimateur de la variance. On rappelle la définition de la variance : $\mathbb{V}(X) := \mathbb{E}((X - \mathbb{E}(X))^2)$. La définition naïve de l'estimateur consiste juste à prendre la moyenne.

Définition 227. L'**estimateur naïf de la variance** est défini par :

$$S_n^2 := \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Cet estimateur naïf est biaisé (mais asymptotiquement non biaisé) :

 **Proposition 228.** L'estimateur de la variance est biaisé :

$$\mathbb{E}(S_n^2) := \frac{n-1}{n} \sigma^2$$

où σ^2 est la variance théorique (i.e. le paramètre inconnu).

DÉMONSTRATION.

Dans la démonstration, on note X pour désigner une variable aléatoire de même loi que chaque X_i , afin d'alléger les notations. On note $\mu = \mathbb{E}X$. Commençons :

$$\begin{aligned}\mathbb{E}(S_n^2) &= \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2\right) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}((X_i - \bar{X}_n)^2) \\ &= \mathbb{E}((X_1 - \bar{X}_n)^2) \\ &= \mathbb{E}(X_1^2 + (\bar{X}_n)^2 - 2X_1\bar{X}_n) \\ &= \mathbb{E}(X^2) + \mathbb{E}((\bar{X}_n)^2) - 2\mathbb{E}(X_1\bar{X}_n) \text{ par linéarité de l'espérance}\end{aligned}$$

Calcul de $\mathbb{E}(X^2)$.

Voici la formule de Koenig-Huygens (proposition 107) : $\sigma^2 := \mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$. On en déduit

$$\mathbb{E}(X^2) = \sigma^2 + \mu^2.$$

Calcul de $\mathbb{E}((\bar{X}_n)^2)$.

De la même façon, la formule de Koenig-Huygens sur $\mathbb{V}(\bar{X})$ donne

$$\begin{aligned}\mathbb{E}((\bar{X}_n)^2) &= \mathbb{V}(\bar{X}) + \mathbb{E}(\bar{X}_n)^2 \\ &= \frac{\sigma^2}{n} + \mathbb{E}(\bar{X}_n)^2 \text{ car les variables } X_i \text{ sont indépendantes, via corollaire 120} \\ &= \frac{\sigma^2}{n} + \mu^2 \text{ par linéarité de l'espérance}\end{aligned}$$

Calcul de $\mathbb{E}(X_1\bar{X}_n)$.

$$\begin{aligned}\mathbb{E}(X_1\bar{X}_n) &= \mathbb{E}\left(\frac{1}{n}[X_1X_1 + X_1X_2 + \dots + X_1X_n]\right) \text{ par définition de } \bar{X}_n \\ &= \frac{1}{n}(\mathbb{E}(X_1^2) + \mathbb{E}(X_1X_2) + \dots + \mathbb{E}(X_1X_n)) \text{ par linéarité de l'espérance} \\ &= \frac{1}{n}(\mathbb{E}(X_1^2) + \mathbb{E}(X_1)\mathbb{E}(X_2) + \dots + \mathbb{E}(X_1)\mathbb{E}(X_n)) \text{ car les } X_i \text{ sont indépendantes entre elles} \\ &= \frac{1}{n}(\mathbb{E}(X_1^2) + (n-1)\mu^2) \\ &= \frac{1}{n}(\sigma^2 + \mu^2 + (n-1)\mu^2) \text{ par la formule de Koenig-Huygens} \\ &= \mu^2 + \frac{\sigma^2}{n}\end{aligned}$$

Bilan. Revenons sur

$$\begin{aligned}\mathbb{E}(S_n^2) &= \mathbb{E}(X^2) + \mathbb{E}((\bar{X}_n)^2) - 2\mathbb{E}(X_1\bar{X}_n) \\ &= (\sigma^2 + \mu^2) + \frac{\sigma^2}{n} + \mu^2 - 2 \times \left(\mu^2 + \frac{\sigma^2}{n}\right) \\ &= \frac{n-1}{n}\sigma^2.\end{aligned}$$

■

On peut le corriger pour obtenir un estimateur non biaisé de la variance, en remplaçant le $\frac{1}{n}$ par $\frac{1}{n-1}$.

Définition 229. *L'estimateur de la variance sans biais est*

$$S_n'^2 := \frac{n}{n-1}S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Proposition 230. *L'estimateur de la variance sans biais est bien sans biais.*

10.4 Maximum de vraisemblance

une des méthodes pour construire un estimateur

10.4.1 Vraisemblance

La vraisemblance mesure de combien les données x_1, \dots, x_n collent à la mesure de probabilité de paramètre θ . L'idée est que l'on veut les probabilités d'avoir chaque donnée x_i grandes, i.e. on veut que $\mathbb{P}(X = x_i; \theta)$ soit grands, pour $i = 1..n$. Ainsi, pour mesurer la grandeur de ces probabilités, on les multiplie. C'est la définition de la vraisemblance.

Définition 231 (vraisemblance pour une loi discrète). Soit θ le paramètre d'un modèle statistique où la loi de X est discrète. Soit x_1, \dots, x_n des données. La vraisemblance des données x_1, \dots, x_n est

$$\mathcal{L}(x_1, \dots, x_n ; \theta) := \prod_{i=1}^n \mathbb{P}(X = x_i; \theta).$$

Exemple 232. Soit θ la probabilité de faire pile (i.e. que $X = 1$). On considère les données :

$$\mathcal{L}(x_1, \dots, x_n ; \theta) := \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{\sum_{i=1}^n (1 - x_i)}.$$

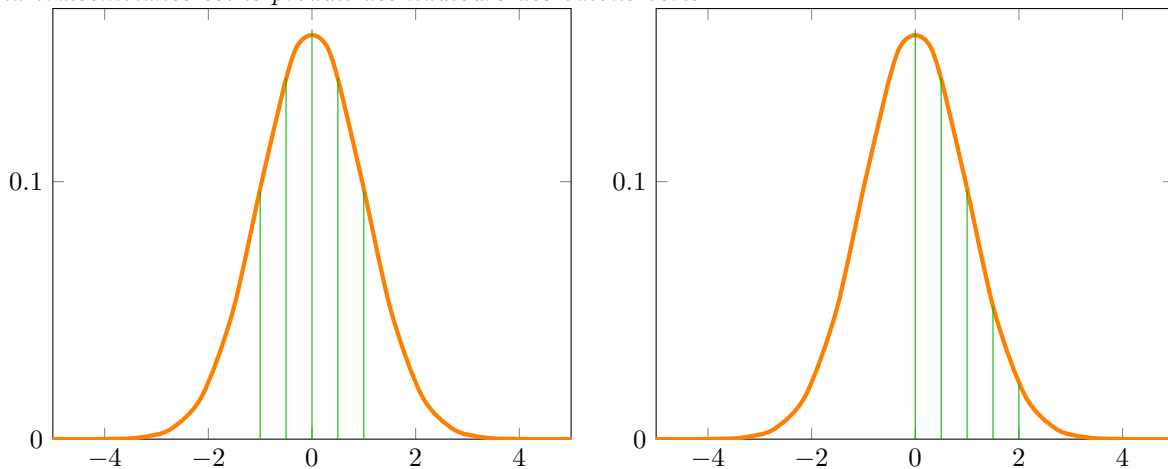
Pour une loi continue, les probabilités $\mathbb{P}(X = x_i; \theta)$ sont nulles (bah c'est vraiment dur d'avoir pile poil la donnée 42!). Du coup, on prend la fonction de densité $p(\bullet; \theta)$ au lieu de prendre la mesure de probabilité $\mathbb{P}(X = \bullet; \theta)$.

Définition 233 (vraisemblance pour une loi continue).

$$\mathcal{L}(x_1, \dots, x_n ; \theta) := \prod_{i=1}^n p(x_i; \theta).$$

où $p(\bullet; \theta)$ est la fonction de densité pour le paramètre θ .

Exemple 234. On considère un modèle statistique où la variable X est de loi $\mathcal{N}(\theta, 1)$. Autrement dit, le paramètre θ est la moyenne de la loi normale. La densité $p(\bullet; \theta) = p_{\mathcal{N}(\theta, 1)}$. Voici le dessin de la fonction de densité pour $\mathcal{N}(0, 1)$. On considère les données $(x_1, x_2, x_3, x_4, x_5) = (-1, -0.5, 0, 0.5, 1)$ et $(x_1, x_2, x_3, x_4, x_5) = (0, 0.5, 1, 1.5, 2)$. A chaque fois, la vraisemblance est le produit des hauteurs des bâtons verts.



10.4.2 Maximiser la vraisemblance

On cherche à maximiser la vraisemblance, i.e. à trouver la valeur du paramètre θ pour laquelle la vraisemblance est maximale. On rappelle que les données x_1, \dots, x_n , elles, sont fixées.

Estimation

entrée : une suite de données x_1, \dots, x_n ;

sortie : une valeur de θ pour laquelle $\mathcal{L}(x_1, \dots, x_n ; \theta)$ est maximale

Cette valeur des paramètres est appelée estimateur du maximum de vraisemblance.

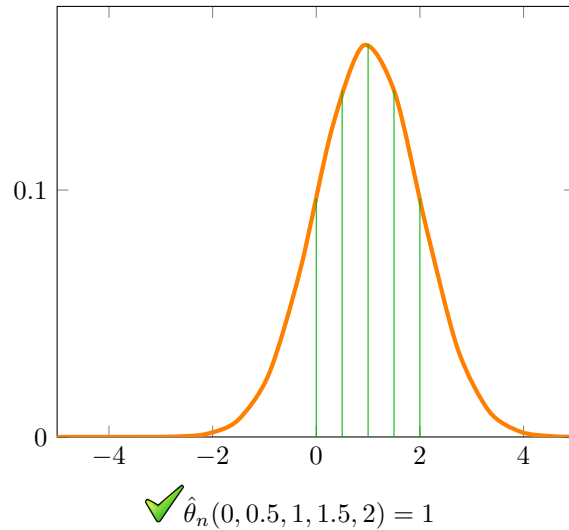
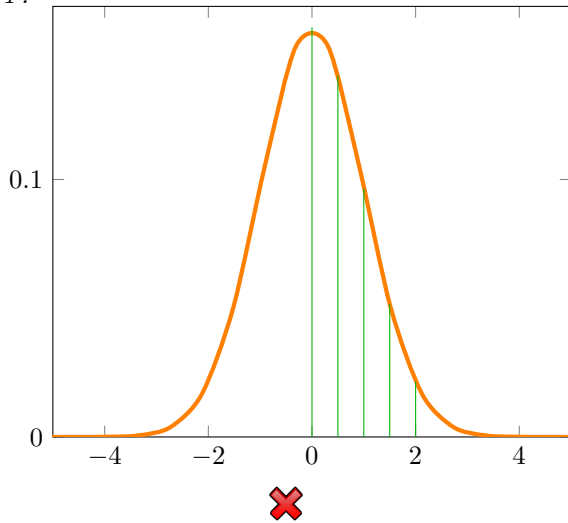
Définition 235. Un *estimateur de maximum de vraisemblance* est une fonction $\hat{\theta}_n$ telle que

$$\hat{\theta}_n(x_1, \dots, x_n) \in \operatorname{argmax}_{\theta \in \Theta} \mathcal{L}(x_1, \dots, x_n ; \theta)$$

Exemple 236. On cherche à trouver l'espérance θ qui fait coller le plus une gaussienne $\mathcal{N}(\theta, 1)$ aux données

$$0, 0.5, 1, 1.5, 2.$$

Cela revient donc à trouver θ qui maximise $\mathcal{L}(0, 0.5, 1, 1.5, 2 ; \theta)$. Bah... la réponse c'est de prendre une espérance θ de 1 !



10.4.3 Exemples

Complètement dingue, mais le maximum de vraisemblance donne des formules plutôt naturelles pour la moyenne et la variance.

Théorème 237. Soit $X \sim \mathcal{B}(\theta)$. L'estimateur par maximum de vraisemblance de θ est



$$(x_1, \dots, x_n) \mapsto \frac{1}{n} \sum_{i=1}^n x_i.$$

DÉMONSTRATION.

$$\mathcal{L}(x_1, \dots, x_n ; \theta) = \prod_{i=1}^n \mathbb{P}(X = x_i ; \theta) = \dots = \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i}$$

Il est difficile de trouver le max de cette fonction selon θ . On passe au log.

$$\log \mathcal{L}(x_1, \dots, x_n ; \theta) = \log \theta \sum x_i + \log(1 - \theta)(n - \sum x_i)$$

Pour trouver le max, on dérive par rapport à θ :

$$\frac{\partial \log \mathcal{L}(x_1, \dots, x_n ; \theta)}{\partial \theta} = \frac{\sum x_i}{\theta} - \frac{n - \sum x_i}{1 - \theta}$$

En faisant $\frac{\partial \log \mathcal{L}(x_1, \dots, x_n ; \theta)}{\partial \theta} = 0$, on trouve

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

■

Théorème 238. Soit $X \sim \mathcal{N}(\mu, \sigma^2)$. L'estimateur par maximum de vraisemblance de (μ, σ) est donné par :



$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

DÉMONSTRATION.

$$\begin{aligned}\mathcal{L}(x_1, \dots, x_n ; \mu, \sigma) &= \prod_{i=1}^n p_{\mathcal{N}(\mu, \sigma^2)}(x_i) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2}\end{aligned}$$

$$\log \mathcal{L}(x_1, \dots, x_n ; \mu, \sigma) = n \log(\sqrt{2\pi}\sigma) - \sum_{i=1}^n \frac{1}{2} \left(\frac{x_i - \mu}{\sigma} \right)^2$$

Dérivée nulle selon μ donne $\hat{\mu}$ du théorème. Dérivée nulle selon σ donne $\hat{\sigma}$ du théorème.

■

10.5 Méthode des moments (*)

La méthode des moments est une alternative pour construire des estimateurs. Je trouve qu'elle est moins utilisée que celle du maximum de vraisemblance. Mais pour la culture, autant la présenter.

Définition 239. Soit $k \in \mathbb{N}^*$. Le **moment d'ordre k** de la variable aléatoire X est $\mathbb{E}(X^k)$.

Exemple 240. Si $X \sim \mathcal{E}(\theta)$, alors $\mathbb{E}(X^k) = \frac{k!}{\theta^k}$.

Par la loi des grands nombres (théorème 162 page 54), étant donné $(X_n)_{n \in \mathbb{N}^*}$ iid, on a

$$\frac{1}{n} \sum_{i=1}^n X_i^k \xrightarrow[n \rightarrow +\infty]{p.s.} \mathbb{E}(X_1^k)$$

Définition 241. On considère une expression mathématique de θ en fonction de $\mathbb{E}(X)$, $\mathbb{E}(X^2)$, $\mathbb{E}(X^3)$, etc. Un **estimateur par la méthode des moments** de θ est l'expression mathématique de θ dans laquelle on a remplacé $\mathbb{E}(X^k)$ par $\frac{1}{n} \sum_{i=1}^n X_i^k$.

Exemple 242. On considère un modèle statistique de loi exponentielle $\mathcal{E}(\theta)$ où θ est le paramètre inconnu. Si $X \sim \mathcal{E}(\theta)$, alors $\mathbb{E}(X) = \frac{1}{\theta}$. Voici donc l'expression de θ :

$$\theta = \frac{1}{\mathbb{E}(X)}.$$

Voici un estimateur par la méthode des moments :

$$\hat{\theta}_n := \frac{n}{\sum_{i=1}^n X_i}.$$

Il n'y a pas unicité de l'estimateur par la méthode des moments. Il y a en fait un estimateur par expression mathématique de θ .

Exemple 243. On a aussi $\mathbb{E}(X^2) = \frac{2}{\theta^2}$. On en déduit donc :

$$\theta = \sqrt{\frac{2}{\mathbb{E}(X^2)}}.$$

Voici un **autre** estimateur par la méthode des moments :

$$\hat{\theta}_n = \sqrt{\frac{2n}{\sum_{i=1}^n X_i^2}}.$$

Exercices

Démonstration du biais de l'estimateur de la variance.

Calcul du maximum de vraisemblance pour la moyenne d'une loi de Bernoulli.

Montrer que l'estimateur du modèle statistique exponentielle par la méthode des moments est biaisé.

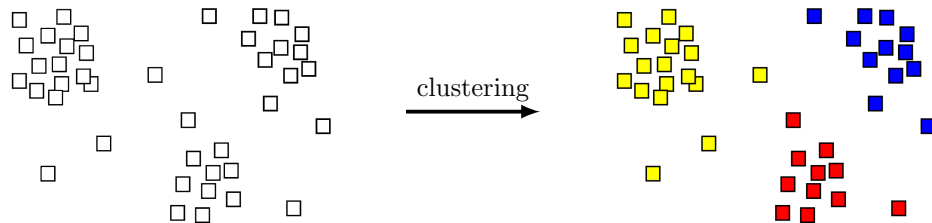
Chapitre 11

Clustering

Clustering

entrée : des données $x_1, \dots, x_n \in \mathbb{R}^p$;

sortie : des clusters, i.e. une ‘bonne’ partition des données



Dans ce chapitre, nous allons voir le problème de clustering comme un problème d'estimations avec une loi mélange : on estime la position des clusters (des moyennes), leurs formes (matrice de covariance), leur importance.

11.1 Applications

Regrouper des acheteurs en fonction des produits déjà achetés.

→ on peut alors recommander des produits adéquats aux acheteurs d'un même cluster

Reconnaître des communautés dans un réseau social

→ proposer des amis

Segmenter une image en plusieurs parties

11.2 Clusters

Définition 244 (partition en clusters). *Il s'agit d'un partitionnement*

$$\{x_1, \dots, x_n\} = \bigcup_{k=1}^K C_k$$

avec $C_k \neq \emptyset$.

Définition 245 (cluster). *Chaque C_k s'appelle un cluster.*

Définition 246 (centroïde). *Le centroïde μ_{C_k} du cluster C_k est le point*

$$\mu_{C_k} := \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i.$$

Le centroïde μ_{C_k} est l'équibarycentre des données du cluster C_k .

11.3 Algorithme des k -means

11.3.1 Problème

Clustering résolu par l'algorithme des k -means

entrée : des données $x_1, \dots, x_n \in \mathbb{R}^p$;

sortie : une partition en k clusters C_1, \dots, C_k qui minimise

$$\sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - \mu_{C_k}\|_2^2.$$

où μ_{C_k} est le centroïde de C_k pour tout $k = 1..K$ et où $\|\cdot\|_2^2$ est le carré de la norme euclidienne.

11.3.2 Algorithme de Lloyd

```

fonction Lloyd( $x_1, \dots, x_n$ )
    choisir  $\mu_1, \dots, \mu_K \in \{x_1, \dots, x_n\}$  distincts
    pendant que la partition change
        initialiser  $C_k := \{x_i \mid \|x_i - \mu_k\|_2 \leq \|x_i - \mu_{k'}\|_2 \text{ pour tout } k' \neq k\}$  pour tout  $k = 1..K$ 
        pour  $i = 1..n$ 
            | mettre  $x_i$  dans  $C_k$  où  $k$  minimise  $\|x_i - \mu_k\|_2$ 
        pour  $k = 1..K$ 
            | on recalcule les centres :  $\mu_k := \mu_{C_k}$ 

```

On commence par des centres qui sont des points distincts parmi x_1, \dots, x_n , choisi au pif. Puis on applique les opérations suivantes : construire les clusters autour de ces “centres”. On assigne ensuite les centroïdes de ces clusters comme nouveaux “centres”. On continue ces opérations tant que la partition change.

Proposition 247. *L'algorithme de Lloyd est en $O(npKt)$ où n = nombre de données, p = dimension des données, K = nombre de clusters, t = nombre d'itérations.*

▲ L'algorithme peut tomber dans un minimum local.

11.3.3 Optimisations

11.4 Clustering par modèle de mélange gaussien

On le verra techniquement plus tard. Mais l'algorithme des k -moyennes vu jusqu'ici ne tient pas compte de la forme des clusters (clusters allongés etc.). Ici, nous allons explicitement introduire la forme des clusters via les matrices de covariances.

Clustering par modèle de mélange gaussien

entrée : $x_1, \dots, x_n \in \mathbb{R}^p$

sortie : des valeurs pour les paramètres $\theta := (w_1, \mu_1, \Sigma_1, \dots, w_K, \mu_K, \Sigma_K)$ modélisant un mélange gaussien

$$w_1 \mathcal{N}(\mu_1, \Sigma_1) + \dots + w_K \mathcal{N}(\mu_K, \Sigma_K)$$

tel que le maximum de vraisemblance par rapport aux données soit maximisée.

Typage

entrée :	p $x_i \in \mathbb{R}^p$	nombre de dimensions pour une donnée i -ème donnée
sortie : paramètre θ à calculer	$w_k \in [0, 1]$ $\mu_k \in \mathbb{R}^p$ $\Sigma_k \in \mathbb{R}^{p \times p}$	poids du cluster k moyenne du cluster k matrice de covariance du cluster k

Dans le problème, w_k est le poids du cluster k , i.e. sa “force” : bref c’est la probabilité qu’une donnée appartienne au cluster k . Le point μ_k est le centre du cluster k . La matrice de covariance Σ_k , elle, donne la forme des clusters (clusters allongés, rétrécis, et dans quelles directions).

11.4.1 Maximum de vraisemblance

Le maximum de vraisemblance des données x_1, \dots, x_n pour la loi mélange gaussienne de densité p est

Déf. 233

$$\begin{aligned}\mathcal{L}(x_1, \dots, x_n ; \theta) &:= \prod_{i=1}^n p(x_i ; \theta) \\ &= \prod_{i=1}^n \sum_{k=1}^K w_k p_{\mathcal{N}(\mu_k, \Sigma_k)}(x_i)\end{aligned}$$

où $p_{\mathcal{N}(\mu_k, \Sigma_k)}$ est la fonction de densité de la loi normale $\mathcal{N}(\mu_k, \Sigma_k)$.

11.4.2 Difficulté de calcul

La log-vraisemblance est

$$\log \mathcal{L}(x_1, \dots, x_n ; \theta) := \sum_{i=1}^n \log \sum_{k=1}^K w_k p_{\mathcal{N}(\mu_k, \Sigma_k)}(x_i)$$

Trouver les paramètres θ qui maximisent la log-vraisemblance semble vraiment compliqué. Le $\log \sum_{k=1}^K$ est gênant et fait que la fonction $\theta \mapsto \log \mathcal{L}(x_1, \dots, x_n ; \theta)$ n'est pas convexe.

11.4.3 Variables latentes

Au lieu de maximiser directement le maximum de vraisemblance (ou de manière équivalente la log-vraisemblance), nous allons faire comme si l'on connaît les numéros des clusters auxquels appartiennent les données x_1, \dots, x_n . On ne peut pas connaître ces numéros de clusters : ce sont des **variables latentes**.

Définition 248. Une **variable latente** est une variable que l'on observe pas directement mais qui intervient dans le modèle.

Ici, si X est la variable aléatoire des données, la variable latente est Z , qui donne le numéro du cluster entre 1 et K de la donnée X . La variable Z suit une loi discrète sur $\{1, \dots, K\}$ avec $\mathbb{P}(Z = k) = w_k$.

La vraisemblance des données x_1, \dots, x_n , munies des valeurs z_1, \dots, z_n de la variable Z est

$$\begin{aligned}\mathcal{L}((x_1, z_1), \dots, (x_n, z_n) ; \theta) &:= \prod_{i=1}^n \mathbb{P}(Z = z_i \text{ et } X = x_i) \\ &= \prod_{i=1}^n \mathbb{P}(Z = z_i) p_{\mathcal{N}(\mu_{z_i}, \Sigma_{z_i})}(x_i) \\ &= \prod_{i=1}^n w_{z_i} p_{\mathcal{N}(\mu_{z_i}, \Sigma_{z_i})}(x_i).\end{aligned}$$

La log-vraisemblance des données x_1, \dots, x_n munies des valeurs z_1, \dots, z_n de la variable Z s'obtient en passant au log :

$$\log \mathcal{L}((x_1, z_1), \dots, (x_n, z_n) ; \theta) := \sum_{i=1}^n \log w_{z_i} + \log p_{\mathcal{N}(\mu_{z_i}, \Sigma_{z_i})}(x_i).$$

11.4.4 Calcul de l'espérance par rapport aux variables latentes

Maintenant, dans l'expression précédent les variables z_1, \dots, z_n sont latentes. Elles ne sont pas accessibles. Du coup, même si elles sont rangés du côté des données, nous allons les considérer comme aléatoires. On introduit Z_1, \dots, Z_n les numéros du cluster respectivement des données x_1, \dots, x_n . On écrit alors :

$$\log \mathcal{L}((x_1, Z_1), \dots, (x_n, Z_n) ; \theta) := \sum_{i=1}^n \log w_{Z_i} + \log p_{\mathcal{N}(\mu_{Z_i}, \Sigma_{Z_i})}(x_i).$$

On écrit maintenant l'espérance de la log vraisemblance . Le paramètre θ est fixé, l'espérance est prise sur les variables latentes aléatoires Z_1, \dots, Z_n .

$$\begin{aligned}
\mathbb{E}(\log \mathcal{L}((x_1, Z_1), \dots, (x_n, Z_n)) ; \theta) &= \mathbb{E}\left(\sum_{i=1}^n \log w_{Z_i} + \log p_{\mathcal{N}(\mu_{Z_i}, \Sigma_{Z_i})}(x_i)\right) \\
&= \sum_{i=1}^n \sum_{k=1}^K \mathbb{P}(Z = k \mid X = x_i) (\log w_k + \log p_{\mathcal{N}(\mu_k, \Sigma_k)}(x_i)).
\end{aligned}$$

11.4.5 Calcul de $\mathbb{P}(Z = k \mid X = x_i)$

La probabilité $\mathbb{P}(Z = k \mid X = x_i)$ est celle qu'un point x_i soit dans le cluster numéro k . Cette probabilité conditionnelle est mal défini car l'événement $X = x_i$ est de mesure nulle. En effet, il est improbable que X vaille exactement x_i !

Pourtant cette probabilité conditionnelle $\mathbb{P}(Z = k \mid X = x_i)$ a l'air d'avoir du sens. Essayons d'écrire la formule de Bayes :

$$\mathbb{P}(Z = k \mid X = x_i) = \frac{\mathbb{P}(Z = k) \mathbb{P}(X = x_i \mid Z = k)}{\mathbb{P}(X = x_i)}$$

Cela ne fonctionne pas, mais donne l'idée de définir la probabilité conditionnelle comme une limite.

$$\begin{aligned}
\mathbb{P}(Z = k \mid X = x_i) &= \lim_{\epsilon \rightarrow 0} \mathbb{P}(Z = k \mid X \in [x_i \pm \epsilon]) \\
&= \lim_{\epsilon \rightarrow 0} \frac{\mathbb{P}(Z = k) \mathbb{P}(X \in [x_i \pm \epsilon] \mid Z = k)}{\mathbb{P}(X \in [x_i \pm \epsilon])} \\
&= \lim_{\epsilon \rightarrow 0} \frac{\mathbb{P}(Z = k) \int_{[x_i \pm \epsilon]} p_{\mathcal{N}(\mu_k, \Sigma_k)}(x_i)(x) dx}{\int_{[x_i \pm \epsilon]} p_{\text{mélange}}(x) dx} \\
&= \frac{\mathbb{P}(Z = k) p_{\mathcal{N}(\mu_k, \Sigma_k)}(x_i)(x_i)}{p_{\text{mélange}}(x_i)} \\
&= \frac{w_k p_{\mathcal{N}(\mu_k, \Sigma_k)}(x_i)}{\sum_{k=1}^K w_k p_{\mathcal{N}(\mu_k, \Sigma_k)}(x_i)}
\end{aligned}$$

Bref, c'est une version continue de la formule de Bayes, qui nous permet d'estimer $\mathbb{P}(Z = k \mid X = x_i)$ pour les paramètres courants de θ . On note $\tau_{k,i}$ la valeur calculée de $\mathbb{P}(Z = k \mid X = x_i)$ pour ces valeurs des paramètres θ là.

11.4.6 Estimation des paramètres maximisant cette espérance

Il s'agit maintenant de calculer θ qui maximise

$$\sum_{i=1}^n \sum_{k=1}^K \tau_{k,i} (\log w_k + \log p_{\mathcal{N}(\mu_k, \Sigma_k)}(x_i)).$$

On notera que les paramètres θ n'apparaissent plus dans les $\tau_{k,i}$ qui sont considéré comme des nombres, mais apparaissent dans l'expression $\log w_k + \log p_{\mathcal{N}(\mu_k, \Sigma_k)}(x_i)$.

11.4.7 Pseudo-code

L'algorithme EM consiste à itérer le calcul de l'espérance (espérance sur les variables latentes, θ fixé) puis la mise à jour de θ avec les paramètres qui maximise l'espérance.

```

fonction algorithmeEM( $x_1, \dots, x_n$ )
  initialiser  $\theta$  au pif
  pendant que on pense que l'algorithme n'a pas fini
    — étape E :
      — estimation des probabilités  $\mathbb{P}(Z = k \mid X = x_i)$  en fonction de  $\theta$  fixé
      — calcul de l'expression formelle  $\mathbb{E}(\log \mathcal{L}(x_1, Z_1, \dots, x_n, Z_n ; \theta))$  où les
        paramètres  $\theta$  sont des variables, sauf dans  $\mathbb{P}(Z = k \mid X = x_i)$  qui ont
        été estimé
    — étape M :  $\theta :=$  valeurs de  $\theta$  qui maximise  $\mathbb{E}(\log \mathcal{L}(x_1, Z_1, \dots, x_n, Z_n ; \theta))$ 

```

11.5 Relation avec K-moyennes

Théorème 249. *La méthode des K-moyennes correspond au clustering de mélange de gaussiennes où on sait que les poids sont de $\frac{1}{K}$ et les matrices de covariances sont égales à l'identité. Autrement dit, le modèle paramétrée est :*

$$\frac{1}{K}\mathcal{N}(\mu_1, Id) + \cdots + \frac{1}{K}\mathcal{N}(\mu_K, Id)$$

où le paramètre inconnu est $\theta := (\mu_1, \dots, \mu_K)$.

Chapitre 12

Lois du χ^2

12.1 Définition

On note $\chi^2(k)$ la loi de carrés de k variables aléatoires iid de loi normale centrée. C'est typiquement la loi d'une **somme d'erreurs quadratiques indépendantes**.

Définition 250. La **loi du χ^2 à k degrés de liberté**, notée $\chi^2(k)$, est celle de $\sum_{i=1}^k X_i^2$ où X_1, \dots, X_k sont iid et suivent $\mathcal{N}(0, 1)$.

12.2 Propriétés

Fonction Gamma

La fonction Gamma généralise la factorielle $n!$ à tout réel positif. Elle intervient dans l'expression de la fonction de densité de la loi du χ^2 .

Définition 251 (fonction Gamma). Pour tout nombre $x > 0$:

$$\Gamma : z \mapsto \int_0^{+\infty} t^{z-1} e^{-t} dt$$

Proposition 252. On a pour tout $x > 0$, $\Gamma(z+1) = z \Gamma(z)$.

Proposition 253. Pour tout entier n , $\Gamma(n+1) = n!$.

Définition 254 (Fonction Gamma incomplète).

$$\gamma(a, x) = \int_0^x t^{a-1} e^{-t} dt.$$

Proposition 255. La fonction de densité de $\chi^2(k)$ est pour tout $x > 0$:

$$f_X(x; k) = \frac{1}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$$

Proposition 256. La fonction de répartition de $\chi^2(k)$ est pour tout $x > 0$:

$$F(x; k) = \frac{\gamma(\frac{k}{2}, \frac{x}{2})}{\Gamma(\frac{k}{2})}$$

12.3 Théorème de Cochran

Théorème 257 (de Cochran). Soit $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \sim \mathcal{N}(0, Id)$. Soit F un sous-espace vectoriel de \mathbb{R}^n .

Soit p_F (p_{F^\perp}) la projection orthogonale sur F (resp. F^\perp). Alors

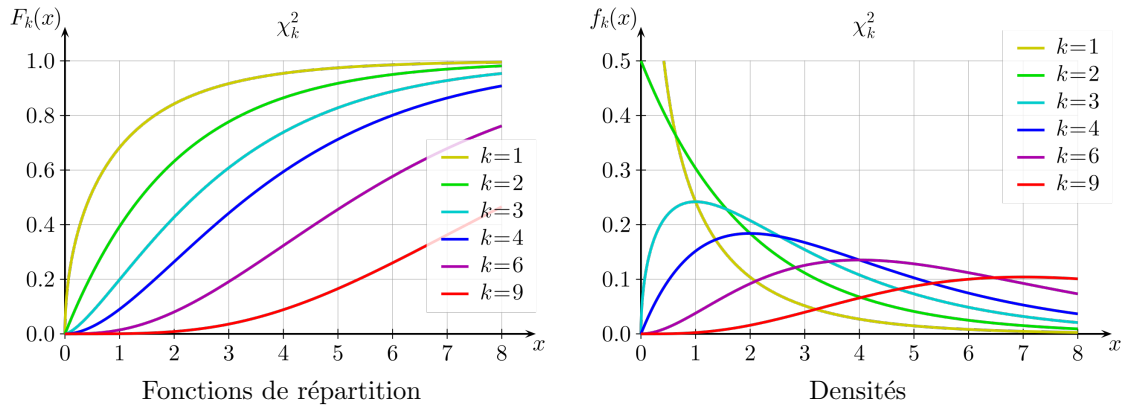


FIGURE 12.1 – Fonctions de répartition et fonctions de densité de la loi du χ^2 à k degrés de liberté.

1. $p_F X \sim \mathcal{N}(0, p_F)$ et $p_{F^\perp} X \sim \mathcal{N}(0, p_{F^\perp})$
2. $p_F X$ et $p_{F^\perp} X$ sont indépendantes (cf. définition 39 page 23)
3. $\|p_F X\|_2^2 \sim \chi^2(\dim F)$ et $\|p_{F^\perp} X\|_2^2 \sim \chi^2(n - \dim F)$

Le théorème de Cochran est l'analogue du théorème de Pythagore.

Chapitre 13

Intervalles de confiance

On peut estimer l'espérance ou la variance à partir des données. Mais a-t-on fait assez de mesures ? Comment quantifier la qualité de notre estimation ?

13.1 Mon premier exemple

On donne ici l'exemple d'un sondage que vous avez déjà sans doute vu au lycée.

Exemple 258 (estimation d'une proportion dans une population). *Une élection va avoir lieu opposant deux candidats, le candidat A et le candidat B. Un sondage est effectué sur $n = 1024$ électeurs. Les résultats du sondage donne 390 électeurs prêts à voter pour A. Que peut-on dire la réelle proportion d'électeurs favorables à A dans l'ensemble du corps électoral ?*

Soit $X_i = \begin{cases} 1 & \text{si la } i\text{-ème personne vote pour A} \\ 0 & \text{sinon} \end{cases}$.

Soit θ la proportion de votant pour A dans l'ensemble du corps électoral. On suppose que $X_1, \dots, X_n \sim \mathcal{B}(\theta)$ iid. On suppose que ces variables sont indépendantes (on peut representer le sondage à la même personne). Le paramètre θ est inconnu. On peut l'estimer avec la moyenne \bar{X}_n .

Notation 259. On note $[a \pm b]$ l'intervalle $[a - b, a + b]$.

Au lycée, on énonce ce résultat approximatif :

$$\blacktriangle \quad \mathbb{P}(\theta \in [\bar{X}_n \pm 1.96 \frac{\sqrt{\bar{X}_n(1-\bar{X}_n)}}{\sqrt{n}}]) = 0.95.$$

Patience ! Nous allons le ré-énoncer plus tard (voir théorème 277). Mais c'est un bon point de départ pour comprendre l'intuition.

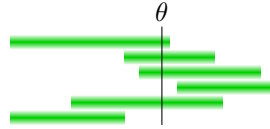
L'intervalle $[\bar{X}_n \pm 1.96 \frac{\sqrt{\bar{X}_n(1-\bar{X}_n)}}{\sqrt{n}}]$ est l'intervalle de confiance. L'intervalle est **aléatoire** car il dépend de \bar{X}_n qui aléatoire. On remarque que plus n grandit, plus $\frac{1}{\sqrt{n}}$ est petit et donc plus l'intervalle se resserre. C'est normal : plus on a de données, plus on est confiant.

Dans la suite on va :

1. Donner la définition formelle d'un intervalle de confiance
2. Comprendre d'où vient le 1.96 : c'est la notion de quantile.
3. Étudier la recette de cuisine pour construire un intervalle de confiance.
4. Étudier la notion d'intervalle de confiance asymptotique (car le cas d'un sondage est en fait plus compliqué et rentre dans ce cadre).

13.2 Définition

Un intervalle de confiance est un intervalle aléatoire qui contient la valeur réelle du paramètre θ avec une forte probabilité. A chaque fois que l'on fait un sondage, il y a une réalisation de l'intervalle aléatoire. Et on a une forte chance que θ se trouve dans l'intervalle.



Définition 260. Un **intervalle de confiance** pour le paramètre θ de **niveau de confiance** $1 - \alpha \in]0, 1[$ est un intervalle aléatoire $I : \Omega \rightarrow \{\text{Intervalles}\}$ avec pour tout $\theta \in \Theta$,

$$\mathbb{P}_\theta(\theta \in I) \geq 1 - \alpha.$$

Parfois, on trouve des définitions avec $\mathbb{P}_\theta(\theta \in I) = 1 - \alpha$.

13.3 Quantiles

Définition 261. Soit X une va réelle. Soit $\alpha \in [0, 1]$. Un **α -quantile** de X est un nombre q_α avec $F_X(q_\alpha) = \alpha$ où F_X est la fonction de répartition de X .

Bref, il y a une probabilité de α que $X \leq q_\alpha$.

Quantiles, déciles, centiles

En pratique, quand on discute avec des statisticiens et statisticiennes, on parle parfois de quantiles particuliers : le médian, les quartiles, déciles, centiles.

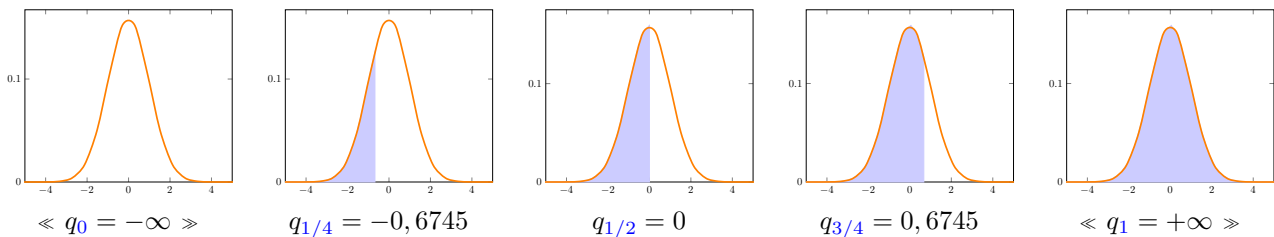
Définition 262.

- Le **médian** est le quantile $q_{0.5}$.
- Les **quartiles** sont les quantiles $q_{0.25}$, $q_{0.5}$, $q_{0.75}$.
- Les **déciles** sont les quantiles $q_{0.1}$, $q_{0.2}$, $q_{0.3}$, $q_{0.4}$, $q_{0.5}$, $q_{0.6}$, $q_{0.7}$, $q_{0.8}$, $q_{0.9}$.
- Les **centiles** sont les quantiles $q_{0.01}$, $q_{0.02}$, $q_{0.03}$, etc.

Attention, un quantile d'ordre α n'est pas forcément unique ! S'il en a plusieurs, certaines personnes disent que le quantile est l'ensemble des nombres x avec $F_X(x) = \mathbb{P}(X \leq x) = \alpha$. Dans nos exemples, F_X est toujours inversible car on va considérer que des variables X à loi continue. Du coup, on a un unique α -quantile de X :

$$q_\alpha = F_X^{-1}(\alpha).$$

Exemple 263 (quantiles vu sur la fonction de densité de la loi normale centrée réduite). Soit $X \sim \mathcal{N}(0, 1)$. On ajoute à chaque fois 1/4 d'eau, et la gravité va vers la gauche. A la surface, on lit à chaque étape le quantile.



Pour une loi normale centrée réduite, le quantile d'ordre $1/2$ vaut 0 , puisque qu'il y a une probabilité de $1/2$ d'être négatif (sous le seuil 0). Le quantile d'ordre $1/4$ vaut $-0,6745$. Le quantile d'ordre $3/4$ vaut $0,6745$.

Comme vu dans l'exemple précédent, nous une loi normale centrée, les quantiles sont symétriques dans le sens suivants. Par exemple, $q_{1/4} = -q_{3/4}$.

Proposition 264. Soit $X \sim \mathcal{N}(0, \sigma)$. Alors pour tout $\alpha \in]0, 1[$ on a $q_{1-\alpha} = -q_\alpha$.

DÉMONSTRATION.



Soit $X \sim \mathcal{N}(0, \sigma)$. Alors la fonction F_X vérifie $F_X(-x) = 1 - F(x)$ pour tout $x \in \mathbb{R}$. Ainsi, on a :

$$\begin{aligned} F_X(q_{1-\alpha}) &= 1 - \alpha \\ &= 1 - F_X(q_\alpha) \\ &= 1 - F_X(-q_{1-\alpha}) \end{aligned}$$

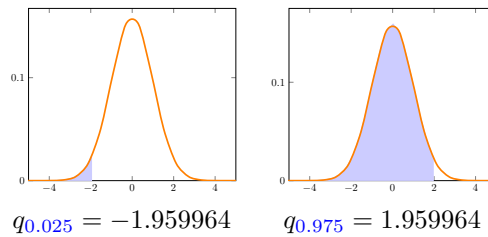
par définition du quantile $q_{1-\alpha}$
 par définition du quantile q_α
 par la propriété de symétrie

Ainsi, on a $F_X(q_\alpha) = F_X(-q_{1-\alpha})$. Comme F_X est injective, on a :

$$q_\alpha = -q_{1-\alpha}.$$

■

Exemple 265 (quantiles symétriques).



Le quantile d'ordre 0.025 est -1.959964. La probabilité d'avoir une valeur sous -1.959964 est 0.025.

13.4 Estimation de la moyenne d'une loi normale avec écart-type connu

Je sais bien que vous mourez d'envie de calculer l'intervalle de confiance pour le cas du sondage. Mais les calculs sont trop compliqués. Entraînons nous sur un cas simple : une loi normale avec écart-type connu.

! Seule la moyenne est inconnue. Autrement dit θ est la moyenne μ à estimer.

Recette de cuisine pour obtenir un intervalle de confiance

Dans la suite, on utilise cette définition.

Définition 266. Une statistique (i.e. une variable aléatoire) est **pivotal** si sa loi est la même quelque soit le paramètre θ .

Voici la recette de cuisine à appliquer pour obtenir un intervalle de confiance.

1. Définir une statistique **pivotal**. L'expression de cette statistique contient le paramètre θ , forcément pour faire que la loi ne dépend pas de θ .
2. Étudier la loi de cette statistique. Généralement, c'est une loi connue déjà, comme $\mathcal{N}(0, 1)$.
3. Établir des quantiles de cette loi. Pareil, la loi est connue, donc pas de soucis.
4. Donner un intervalle de confiance. On reprend l'encadrement de la statistique avec des quantiles que l'on retourne comme une chaussette pour avoir l'intervalle de confiance.

13.4.1 Statistique pivotale

Proposition 267. Soit X_1, \dots, X_n un échantillon aléatoire de loi $\mathcal{N}(\mu, \sigma^2)$. Alors

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

où μ est la moyenne de la loi normale, σ est l'écart-type de la loi normale, \bar{X}_n est la moyenne des X_i .

DÉMONSTRATION.

X_1, \dots, X_n un échantillon aléatoire de loi $\mathcal{N}(\mu, \sigma^2)$

Donc

$$X_1 + \dots + X_n \sim \mathcal{N}(n\mu, n\sigma^2).$$

En divisant par n , on divise la variance par n^2 , on a donc :

$$\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n) \sim \mathcal{N}\left(\mu, \frac{1}{n^2}n\sigma^2\right) = \mathcal{N}\left(\mu, \frac{1}{n}\sigma^2\right).$$

On applique ensuite le changement vers loi normale réduite (proposition 71 page 29) que l'on réexplique ici :

$$\bar{X}_n - \mu \sim \mathcal{N}\left(0, \frac{1}{n^2}n\sigma^2\right)$$

c'est-à-dire

$$\bar{X}_n - \mu \sim \mathcal{N}(0, \frac{1}{n}\sigma^2)$$

Ainsi en divisant par l'écart-type σ/\sqrt{n} on obtient :

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

■

13.4.2 Intervalle de confiance

Théorème 268 (intervalle de confiance pour la moyenne μ quand écart-type σ est connu). *Soit X_1, \dots, X_n un échantillon aléatoire de loi $\mathcal{N}(\mu, \sigma^2)$. Alors voici un intervalle de confiance pour la moyenne μ de niveau $1 - \alpha$:*

$$\left[\bar{X}_n \pm q_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de $\mathcal{N}(0, 1)$.

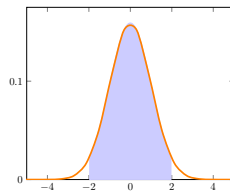
DÉMONSTRATION.

On pose

$$Y_n := \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu).$$

On a vu que c'est une statistique pivotale et qu'elle suit $\mathcal{N}(0, 1)$. Ainsi :

$$\begin{aligned} \mathbb{P}_\theta(Y_n \in [0 \pm q_{\alpha/2}]) &= 1 - \mathbb{P}_\theta(Y_n \leq q_{\alpha/2}) - \mathbb{P}_\theta(Y_n \geq q_{1-\alpha/2}) \\ &= 1 - 2 \times \mathbb{P}_\theta(Y_n \leq q_{\alpha/2}) \\ &= 1 - 2\alpha/2 = 1 - \alpha \end{aligned}$$



Et maintenant le retournement de chaussettes :

$$Y_n \in [0 \pm q_{\alpha/2}] \text{ iff } \bar{X}_n - \mu \in \left[0 \pm \frac{\sigma}{\sqrt{n}} q_{\alpha/2} \right] \text{ iff } \mu \in \left[\bar{X}_n \pm \frac{\sigma}{\sqrt{n}} q_{\alpha/2} \right]$$

Ainsi :

$$\mathbb{P}_\theta(\mu \in \left[\bar{X}_n \pm q_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]) = 1 - \alpha$$

■

Exemple 269.

$$\mathbb{P}_\theta(\mu \in \left[\bar{X}_n \pm 1.96 \frac{\sigma}{\sqrt{n}} \right]) = 1 - 0.05$$

$\left[\bar{X}_n \pm 1.96 \frac{\sigma}{\sqrt{n}} \right]$ est un intervalle de confiance pour l'espérance inconnue μ de niveau de confiance 0.95.

13.5 Estimation de la moyenne d'une loi normale avec écart-type inconnu

! Le paramètre θ est maintenant (μ, σ) où μ est la moyenne inconnue à estimer, et σ est l'écart-type inconnu, lui aussi, mais il n'est pas à estimer.

13.5.1 Se débarrasser de σ

L'intervalle de confiance que l'on vient de voir, dans le théorème 268 est

$$\left[\bar{X}_n \pm q_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right].$$

Malheureusement, σ y apparaît. Comme nous ne connaissons pas la valeur de σ , nous ne pouvons pas calculer l'intervalle de confiance à partir des seules données X_1, \dots, X_n et de n . Bref, σ est un **paramètre nuisible**, comme les cafards que l'on souhaite voir disparaître de sa cuisine.

Nous souhaitons un intervalle de confiance où le terme σ n'apparaît plus.

Nous allons reprendre la même démarche que précédemment. Nous avons toujours

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

qui suit la loi normale centrée réduite. Mais utiliser cette statistique laisse σ dans l'expression de l'intervalle de confiance.

L'idée est de remplacer la variance σ^2 par l'**estimateur de la variance sans biais** (cf. définition 229 page 75) :

$$S_n'^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

De manière surprenante, on obtient une statistique pivotale

$$\frac{\sqrt{n}}{S'_n}(\bar{X}_n - \mu) \sim t(n-1)$$

qui suit une **loi de Student** à $n-1$ degrés de liberté, que l'on note $t(n-1)$ voir définition qui suit.

Loi de Student à n degrés de liberté

Définition 270. Soit $Y \sim \mathcal{N}(0, 1)$ et $Z \sim \chi^2(n)$ indépendantes. La **loi de Student à n degrés de liberté** est la loi de

$$Y \sqrt{\frac{n}{Z}} \sim t(n).$$

On l'admet ici mais on peut montrer que $t(n) \xrightarrow[n]{\mathcal{L}} \mathcal{N}(0, 1)$. Cela signifie que pour des valeurs de n très grandes, on peut continuer à utiliser la loi normale centrée réduite plutôt que la loi de Student.

13.5.2 Statistique pivotale

Proposition 271. Soit X_1, \dots, X_n un échantillon aléatoire de loi $\mathcal{N}(\mu, \sigma^2)$. Alors :



$$\frac{\sqrt{n}}{S'_n}(\bar{X}_n - \mu) \sim t(n-1).$$

DÉMONSTRATION.

Bon, même si σ est inconnu, il existe et on le note σ . On a le droit de l'utiliser dans les calculs. Remarquons que

$$\begin{aligned}
\frac{\sqrt{n}}{S'_n}(\bar{X}_n - \mu) &= \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \times \frac{\sigma}{S'_n} \\
&= \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \times \frac{1}{\frac{S'_n}{\sigma}} \\
&= \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \times \frac{1}{\sqrt{\frac{S_n'^2}{\sigma^2}}} \\
&= \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \times \sqrt{\frac{n-1}{\frac{n-1}{\sigma^2} S_n'^2}} \\
&= Y_n \sqrt{\frac{n-1}{Z_n}}
\end{aligned}$$

avec $Y_n := \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)$ et $Z_n := \frac{n-1}{\sigma^2} S_n'^2$.

Il reste maintenant à montrer les lemmes suivants pour conclure qu'on est bien en face d'une loi de Student de degré $n-1$.

Lemme 272. $Y_n := \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \sim \mathcal{N}(0, 1)$.

DÉMONSTRATION.

C'est la proposition 267. ■

Lemme 273. $Z_n \sim \chi^2(n-1)$ et Y_n et Z_n sont indépendantes.

DÉMONSTRATION.

Préparons-nous à appliquer le théorème de Cochran (théorème 257 page 85) sur $X' = (X'_1, \dots, X'_n)$ avec $X'_i = \frac{X_i - \mu}{\sigma}$. On peut bien appliquer le théorème sur ce vecteur gaussien car chaque $X'_i \sim \mathcal{N}(0, 1)$ et sont indépendants. Pour la moyenne on a :

$$\bar{X}'_n = \frac{\bar{X}_n - \mu}{\sigma}.$$

Considérons le vecteur $\vec{f} = \frac{1}{\sqrt{n}} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n$ et $F := \text{Vect}(\vec{f})$. D'après le théorème de Cochran :

- $\|p_{F^\perp} X'\|^2 \sim \chi^2(n - \dim F) = \chi^2(n-1)$
- $p_F X'$ et $p_{F^\perp} X'$ sont indépendants.

On a :

1. On rappelle que la coordonnée sur la projection sur le vecteur \vec{f} est donné par le produit scalaire $\langle X', \vec{f} \rangle$. On

$$\text{a donc } p_F X' = \langle X', \vec{f} \rangle \vec{f} = \begin{pmatrix} \bar{X}'_n \\ \vdots \\ \bar{X}'_n \end{pmatrix} = \bar{X}'_n \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \frac{1}{\sqrt{n}} Y_n \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$2. \ p_{F^\perp} X' = X' - p_F X' = \begin{pmatrix} X'_1 - \bar{X}'_n \\ \vdots \\ X'_n - \bar{X}'_n \end{pmatrix}.$$

On remarque que

1. $Y_n = \sqrt{n} \times$ la première coordonnées de $p_F X'$;
2. $\|p_{F^\perp} X'\|^2 = \sum_i (X'_i - \bar{X}'_n)^2 = \frac{1}{\sigma^2} \sum_i (X_i - \bar{X}_n)^2 = \frac{(n-1)}{\sigma^2} S_n'^2 = Z_n$.

Par le lemme des coalitions (lemme 43 page 24), Y_n et Z_n sont indépendants. ■

■

13.5.3 Intervalle de confiance

Proposition 274 (intervalle de confiance pour la moyenne μ quand écart-type σ est inconnu).

Soit X_1, \dots, X_n un échantillon aléatoire de loi $\mathcal{N}(\mu, \sigma^2)$. Alors voici un intervalle de confiance pour la moyenne μ de niveau $1 - \alpha$:

$$\left[\bar{X}_n \pm q_{1-\alpha/2} \frac{S'_n}{\sqrt{n}} \right]$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi de Student à $n - 1$ degré de liberté, et où S'_n est l'estimateur sans biais (cf. définition 229 page 75) de la variance des X_1, \dots, X_n .

DÉMONSTRATION.

C'est le même squelette que le théorème précédent, avec le même retournement de chaussette.

■

13.6 Intervalle de confiance asymptotique


Dans les sections précédentes, on obtenait des résultats non asymptotiques avec la loi normale. Là, on va appliquer le théorème central limite, et donc, on va avoir des résultats asymptotiques.

Définition 275 (intervalle de confiance asymptotique). Un *intervalle de confiance asymptotique* pour le paramètre θ de niveau de confiance $1 - \alpha \in]0, 1[$ est une suite d'intervalles aléatoires $(I_n)_{n \in \mathbb{N}}$ avec pour tout $\theta \in \Theta$,

$$\mathbb{P}_\theta(\theta \in I_n) \xrightarrow{n \rightarrow +\infty} 1 - \alpha.$$

13.6.1 Moyenne inconnu, écart-type connu

Le théorème suivant est comme le théorème 268 mais pour n'importe quelle loi.

Théorème 276 (intervalle de confiance asymptotique pour la moyenne μ quand écart-type σ est connu). Soit  Soit X_1, \dots, X_n un échantillon aléatoire de même loi de moyenne μ inconnue et d'écart-type σ connu. Alors

$$\left[\bar{X}_n \pm q_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

est un intervalle de confiance asymptotique pour la moyenne μ de niveau $1 - \alpha$, où $q_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de $\mathcal{N}(0, 1)$.

DÉMONSTRATION.

■

13.6.2 Sondage

Théorème 277. Soit $X_1, X_2, \dots \sim \mathcal{B}(\theta)$ iid. Alors



$$\left[\bar{X}_n \pm q_{1-\alpha/2} \frac{\sqrt{\bar{X}_n(1 - \bar{X}_n)}}{\sqrt{n}} \right]$$

est un intervalle de confiance asymptotique pour le paramètre θ de niveau de confiance $1 - \alpha$, où $q_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$.

13.7 Intervalle de confiance pour un quantile

Définition 278. Soit X_1, \dots, X_n un échantillon aléatoire. Leur *ordre statistique* est

$$X_{i_1} \leq \dots \leq X_{i_n}$$

autrement dit les valeurs des X_i dans l'ordre croissant, avec i_1, \dots, i_n une permutation de $\{1, \dots, n\}$.

On notera que dans la définition précédente i_1, \dots, i_n sont des variables aléatoires, (même si on les écrit en minuscule pour ne pas confondre avec des intervalles I_1, \dots, I_n).

Théorème 279. Soit X_1, \dots, X_n un échantillon aléatoire qui suivent la même loi à densité. Soit q_p le p -quantile des X_i . Voici un intervalle de confiance pour q_α de niveau $1 - \alpha$:

$$[X_{i_j}, X_{i_k}]$$

où X_{i_1}, \dots, X_{i_n} est l'ordre statistique de X_1, \dots, X_n et j et k tels que $\mathbb{P}(N_{\mathcal{B}(n,p)} \in \{j, k-1\}) \geq 1 - \alpha$.

DÉMONSTRATION.

Il faut montrer que, sous les conditions du théorème, $\mathbb{P}(q_p \in [X_{i_j}, X_{i_k}]) \geq 1 - \alpha$.

Comme les variables X_i sont à densité, on a

$$\mathbb{P}(q_p \in [X_{i_j}, X_{i_k}[) = \mathbb{P}(q_p \in [X_{i_j}, X_{i_k}]).$$

Allons-y. Par définition de q_p , on a $\mathbb{P}(X_i \leq q_p) = p$.

On pose $N = \#\{i = 1..n \mid X_i \leq q_p\}$, i.e. le nombre de points de l'échantillon à être avant le p -quantile. On a :

$$N = \sum_{i=1}^n 1_{X_i \leq q_p}.$$

Ainsi, la variable N est une somme de n variables indépendantes $1_{X_i \leq q_p}$ qui suivent chacune une loi de Bernoulli de paramètre p . Ainsi, La variable N suit bien une loi binomiale $\mathcal{B}(n, p)$.

On a l'équivalence des événements suivants :

$$\begin{aligned} q_p \in [X_{i_j}, X_{i_k}[\text{ ssi } & \begin{aligned} & \text{il y a au moins } j \text{ variables } X_i \text{ avec } X_i \leq q_p \text{ et} \\ & \text{il y a au plus } k-1 \text{ variables } X_i \text{ avec } X_i \leq q_p \end{aligned} \\ \text{ssi } N \in \{j, k-1\} \end{aligned}$$

Pour résumer :

$$\begin{aligned} \mathbb{P}(q_p \in [X_{i_j}, X_{i_k}]) &= \mathbb{P}([X_{i_j}, X_{i_k}[) \\ &= \mathbb{P}(N \in \{j, k-1\}) \\ &\geq 1 - \alpha \end{aligned}$$

■

Corollaire 280. Pour n grand, voici un intervalle de confiance asymptotique pour q_α de niveau $1 - \alpha$:

$$[X_{i_j}, X_{i_k}]$$

où X_{i_1}, \dots, X_{i_n} est l'ordre statistique et

$$\begin{aligned} j &= \lfloor np - q_{1-\alpha/2} \sqrt{np(1-p)} \rfloor \\ k &= \lceil np + q_{1-\alpha/2} \sqrt{np(1-p)} \rceil + 1 \end{aligned}$$

et $q_{1-\alpha/2}$ est le quantile de la loi normale centrée réduite.

DÉMONSTRATION.

On approxime la loi binomiale $\mathcal{B}(n, p)$ par une loi normale $\mathcal{N}(\mu, \sigma^2)$ avec $\mu = np$ et $\sigma^2 = np(1-p)$.

Ainsi, la condition $\mathbb{P}(N_{\mathcal{B}(n,p)} \in \{j, k-1\}) \geq 1 - \alpha$ devient $\mathbb{P}(N_{\mathcal{N}(\mu, \sigma^2)} \in \{j, k-1\}) \geq 1 - \alpha$. Cette condition est :

$$\mathbb{P}(F_{\mathcal{N}(\mu, \sigma^2)} \leq k-1) - \mathbb{P}(F_{\mathcal{N}(\mu, \sigma^2)} \leq j) \geq 1 - \alpha$$

Pour avoir un traitement 'symétrique' on choisit j et k :

$$\begin{aligned} \mathbb{P}(F_{\mathcal{N}(\mu, \sigma^2)} \leq k-1) &\geq \frac{1}{2} + \frac{1-\alpha}{2} = 1 - \frac{\alpha}{2} \\ \mathbb{P}(F_{\mathcal{N}(\mu, \sigma^2)} \leq j) &\leq \frac{1}{2} - \frac{1-\alpha}{2} = \frac{\alpha}{2} \end{aligned}$$

La dernière condition $\mathbb{P}(F_{\mathcal{N}(\mu, \sigma^2)} \leq j) \leq \frac{\alpha}{2}$ est équivalent à :

$$\mathbb{P}\left(\frac{N_{\mathcal{N}(\mu, \sigma^2)} - \mu}{\sigma} \leq \frac{j - \mu}{\sigma}\right) \leq \frac{\alpha}{2}$$

Ainsi, c'est équivalent à

$$\frac{j - \mu}{\sigma} \leq q_{\frac{\alpha}{2}}$$

autrement dit

$$j \leq q_{\frac{\alpha}{2}} \sigma + \mu$$

ou encore

$$j \leq -q_{1-\frac{\alpha}{2}} \sigma + \mu$$

par symétrie des quantiles. On a donc le premier point pour j . La démonstration pour k est similaire (à vérifier). ■

13.8 Intervalle de prédiction

Définition 281. Soit X_1, \dots, X_n, X_{n+1} des variables aléatoires. Un **intervalle de prédiction** de niveau de confiance $1 - \alpha$ est un intervalle aléatoire $I(X_1, \dots, X_n)$ tel que

$$\mathbb{P}(X_{n+1} \in I(X_1, \dots, X_n)) \geq 1 - \alpha.$$

13.8.1 Cas général iid

Théorème 282. Soit X_1, \dots, X_n, X_{n+1} des variables aléatoires iid, ayant une densité. alors on a :

$$\mathbb{P}(X_{i_j} \leq X_{n+1} \leq X_{i_k}) = \frac{k - j}{n + 1}$$

où X_{i_1}, \dots, X_{i_n} est l'ordre statistique de X_1, \dots, X_n .

DÉMONSTRATION.

Soit F la fonction de répartition de la loi d'un X_i . On pose

$$U_i = F(X_i).$$

La variable U_i suit une loi uniforme sur $[0, 1]$.

Loi bêta et distribution de l'ordre statistique de variables aléatoires iid uniformes sur $[0, 1]$

Définition 283. La loi bêta de paramètre k et $n + 1 - k$ est la loi sur $[0, 1]$ est donnée par la densité

$$f(u) = \frac{n!}{(k-1)!(n-k)!} u^{k-1} (1-u)^{n-k}.$$

Proposition 284. U_{i_k} soit la loi bêta de paramètre k et $n + 1 - k$.

DÉMONSTRATION.

Pour le voir, calculons $\mathbb{P}(U_{i_k} \in [u, u + du])$. L'événement $U_{i_k} \in [u, u + du]$ a lieu lorsque $k - 1$ éléments sont dans $[0, u[$, un élément dans $[u, u + du[$ et $n - k$ éléments dans $]u + du, 1]$. La distribution multinomiale donne le résultat. ■

Proposition 285. $\mathbb{E}(U_{i_k}) = \frac{k}{n+1}$.

On a $\mathbb{E}(U_{i_k}) = \frac{k}{n+1}$.

$$\begin{aligned} \mathbb{P}(U_{i_j} \leq U_{n+1} \leq U_{i_k}) &= \mathbb{E}(U_{i_k} - U_{i_j}) \\ &= \frac{k - j}{n + 1} \end{aligned}$$

à détailler...

par linéarité de l'espérance

■

Corollaire 286. Soit X_1, \dots, X_n, X_{n+1} des variables aléatoires iid, ayant une densité. $[X_{i_{\lfloor (n+1)\alpha/2 \rfloor}}, X_{i_{\lceil (n+1)(1-\alpha/2) \rceil}}]$ est un intervalle de prédiction de niveau de confiance au moins $1 - \alpha$.

DÉMONSTRATION.

Ca correspond à avoir :

$$j = \lfloor (n+1)\alpha/2 \rfloor$$

$$k = \lceil (n+1)(1-\alpha/2) \rceil$$

On a alors :

$$\begin{aligned} \mathbb{P}(U_{i_j} \leq U_{n+1} \leq U_{i_k}) &= \frac{k-j}{n+1} \\ &= \frac{\lceil (n+1)(1-\alpha/2) \rceil - \lfloor (n+1)\alpha/2 \rfloor}{n+1} \\ &= \frac{(n+1)(1-\alpha/2) - (n+1)\alpha/2}{n+1} && \text{car ça se compense gentiment} \\ &= \frac{(n+1)(1-\alpha)}{(n+1)} \\ &= (1-\alpha) \end{aligned}$$

■

13.8.2 Cas normal iid

Théorème 287. Soit X_1, \dots, X_n, X_{n+1} iid et qui suivent toutes $\mu\sigma^2$. Voici un intervalle de prédiction au niveau $1 - \alpha$:

$$\left[\bar{X}_n \pm q_{1-\alpha/2} S'_n \sqrt{1 + \frac{1}{n}} \right]$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi de Student à $n - 1$ degré de liberté, et où S'_n est l'estimateur sans biais (cf. définition 229 page 75) de la variance des X_1, \dots, X_n .

DÉMONSTRATION.



X_{n+1} indépendant de \bar{X}_n .

$X_{n+1} - \bar{X}_n$ est normal de moyenne 0 et de variance

$$\mathbb{V}(X_{n+1}) - \mathbb{V}(\bar{X}_n) = \sigma^2 + \frac{\sigma^2}{n}.$$

Bref, avant on avait $\bar{X}_n - \mu \sim \mathcal{N}(0, \frac{1}{n}\sigma^2)$, maintenant on a :

$$X_{n+1} - \bar{X}_n \sim \mathcal{N}(0, \sigma^2(1 + \frac{1}{n})).$$

De plus, $(n-1)\frac{S_n'^2}{\sigma^2} \sim \chi^2(n-1)$ et est indépendant de $X_{n+1} - \bar{X}_n$.

Ainsi, comme dans la proposition 271 on cherche une statistique qui suit une loi de Student à $n - 1$ degré de liberté de la forme :

$$Y_n \sqrt{\frac{n-1}{Z_n}}$$

avec Y_n qui suit une loi normale centrée, et Z_n qui suit $\chi^2(n-1)$.

Il s'agit de :

$$T = \frac{1}{\sigma\sqrt{1+\frac{1}{n}}}(\bar{X}_n - X_{n+1}) \times \sqrt{\frac{n-1}{\frac{n-1}{\sigma^2}S_n'^2}} = \frac{1}{S'_n\sqrt{1+\frac{1}{n}}}(\bar{X}_n - \mu)$$

En effet, la partie verte est bien de la forme

$$\frac{\text{loi normale centrée}}{\text{écart-type}}$$

donc ça suit bien une loi normale centrée réduite.

En retournant la chaussette, on obtient l'intervalle de prédiction :

$$T \in [0 \pm q_{\alpha/2}] \text{ ssi } \bar{X}_n - X_{n+1} \in \left[0 \pm S'_n \sqrt{1 + \frac{1}{n} q_{\alpha/2}} \right] \text{ ssi } X_{n+1} \in \left[\bar{X}_n \pm S'_n \sqrt{1 + \frac{1}{n} q_{\alpha/2}} \right]$$

■

Exercices

Étudier l'intervalle de confiance pour l'écart type (moyenne connue et inconnue).

Chapitre 14

Généralités sur les tests statistiques

Dans le chapitre précédent, nous avons estimé le paramètre de la distribution en fonction de données. Ici, nous allons prendre des **décisions binaires**.

Exemple 288. *Est-ce que ma pièce de monnaie est pipée ? Est-ce que le médicament est meilleur qu'un placebo ? Est-ce que les gènes de la couleur des yeux et des cheveux sont sur le même chromosome ? Est-ce que la clairvoyance existe ?*

Pour répondre à ces questions de recherche, on met en place des **expériences**.

Exemple 289. *Lancer la pièce plusieurs fois. Noter les résultats.*

Exemple 290. *Prenons une carte avec un côté blanc et un côté noir. Faire toucher la carte à une personne. Aller déposer la carte dans une autre pièce. Déposer la carte sur une table, avec une face quelconque vers le haut. S'assurer que la personne ne voit pas la carte (pas de reflet dans une vitre où je ne sais quoi). Demander à la personne la face visible de la carte (blanc ou noir). Noter si la personne se trompe ou non. Répéter l'expérience.*

L'estimation ne permet pas de bien répondre à ce genre de questions. Même si la pièce de monnaie est non pipée, l'estimateur ne donnera jamais exactement 50% de chances de faire pile. Même si la clairvoyance n'existe pas, l'estimateur ne donnera jamais exactement 50% d'erreurs.

À partir de maintenant, on suppose que l'expérience est conçu de façon à bien répondre à la question. On se fixe une expérience.

14.1 Hypothèses

On considère une hypothèse privilégiée, considérée comme la plus vraisemblable, dont le rejet à tort est le plus préjudiciable. On dit qu'elle est « nulle » car elle est notée H_0 . L'autre hypothèse se note H_1 .

Définition 291 (hypothèse nulle). *L'hypothèse H_0 s'appelle l'hypothèse nulle.*

Définition 292 (hypothèse alternative). *L'hypothèse H_1 s'appelle l'hypothèse alternative.*

On sait que soit H_0 ou H_1 vraie. Souvent H_1 est la négation de H_0 , mais pas toujours. Définir H_1 proprement permet de définir proprement la notion de puissance d'un test etc.

Exemple 293. *Voici des exemples informels d'hypothèses nulles et d'hypothèses alternatives :*

<i>la pièce n'est pas pipée</i>	<i>la pièce est pipée</i>
<i>le dé n'est pas pipé</i>	<i>le dé est pipé</i>
<i>la clairvoyance n'existe pas</i>	<i>la clairvoyance existe</i>
<i>les deux échantillons proviennent de la même loi</i>	<i>non, ils proviennent de lois différentes</i>
<i>Math.random() suit une loi uniforme</i>	<i>Math.random() ne suit pas une loi uniforme</i>
<i>Amélie est innocente</i>	<i>Amélie est coupable</i>

Exemple 294 (est-ce que la pièce est équilibrée ?). *On note p la probabilité de faire pile.*

$$H_0 : p = 1/2$$

$$H_1 : p \neq 1/2$$

la pièce est équilibrée

la pièce est pipée

Exemple 295 (est-ce que la clairvoyance existe ?). *On note p la probabilité de se tromper.*

$$H_0 : p = 1/2$$

$$H_1 : p \neq 1/2$$

la clairvoyance n'existe pas

la clairvoyance existe

On a ici défini la clairvoyance comme une propension à dire la vérité $p < 1/2$ ou à se tromper $p > 1/2$.

14.2 Algorithme

Définition 296 (test). Un **test statistique** est un algorithme qui répond au problème :

- entrée : données x_1, \dots, x_n
- sortie : conserver H_0 , ou rejeter H_0 .

Généralement, un test statistique prend la forme suivante :

```

fonction test( $x_1, \dots, x_n$ )
|   calculer une certaine quantité  $T(x_1, \dots, x_n)$ 
|   si la valeur  $T(x_1, \dots, x_n)$  n'est pas plausible selon  $H_0$  alors
|   |   rejeter  $H_0$ 
|   sinon
|   |   conserver  $H_0$ 

```

14.3 Exemple : pièce équilibrée ou pipée ?

On note p la probabilité que un lancer de la pièce donne pile.

$$\begin{aligned}
 H_0 : & \quad p = 1/2 \text{ (la pièce est équilibrée)} \\
 H_1 : & \quad p \neq 1/2 \quad \text{(la pièce est pipée)}
 \end{aligned}$$

On va concevoir un test pour ça :

- entrée : données x_1, \dots, x_n où $x_i = \begin{cases} 1 & \text{si le } i\text{-ème lancer est pile} \\ 0 & \text{sinon} \end{cases}$
- sortie : conserver H_0 , ou rejeter H_0 .

```

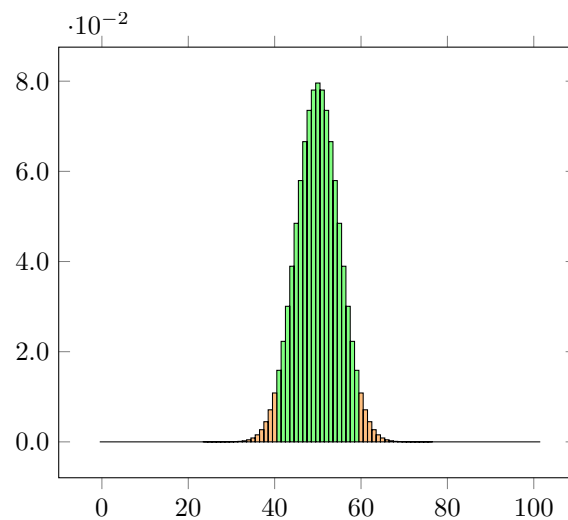
fonction testPiece( $x_1, \dots, x_{100}$ )
|   calculer la statistique  $\hat{N} := \sum_{i=1}^{100} x_i$ 
|   si  $\hat{N} \in [41, 59]$  alors conserver  $H_0$  sinon rejeter  $H_0$ 

```

Supposons H_0 , i.e. $p = 1/2$. Alors $X_1, \dots, X_{100} \sim \mathcal{B}(1/2)$ et sont indépendantes. Alors $N = \sum_{i=1}^{100}$ suit $\mathcal{B}(1/2, 100)$. On a

$$\mathbb{P}(N \in [41, 59]) = 0.95.$$

Proposition 297. Le risque de première espèce *testPiece* est $\alpha = 5\%$.



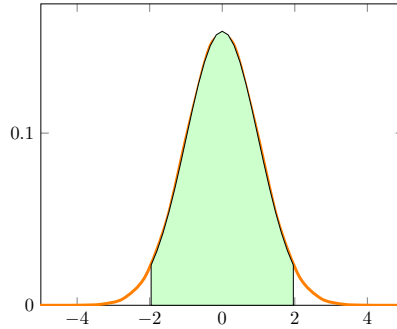
14.4 Intervalle de fluctuation

De manière générale, l'intervalle vert, comme [41, 59] s'appelle un intervalle de fluctuation.

Définition 298. Soit X une va. Un **intervalle de fluctuation** au seuil de p est un intervalle I non aléatoire tel que

$$\mathbb{P}(X \in I) = p.$$

Exemple 299. Soit $X \sim \mathcal{N}(0, 1)$ alors $[-1.96, 1.96]$ est un intervalle de fluctuation au seuil de 95%. Le dessin suivant montre la fonction de densité de $\mathcal{N}(0, 1)$. L'aire coloriée est $\mathbb{P}(X \in [-1.96, 1.96])$ et vaut environ 0.95.



Exemple 300. Soit $X \sim \mathcal{N}(\mu, \sigma^2)$ alors $[\mu \pm 1.96\sigma]$ est un intervalle de fluctuation au seuil de 95%.

Exemple 301. Soit X_1, \dots, X_n qui suivent une loi de Bernoulli d'espérance p et de variance $p(1-p)$. On approxime $\frac{1}{n} \sum_{i=1}^n X_i$ via le TCL (théorème 390) par $\mathcal{N}(p, p(1-p)/n)$. Ainsi, l'intervalle de fluctuation est

$$\left[p \pm 1.96 \sqrt{\frac{p(1-p)}{n}} \right].$$

14.5 Test z

On suppose que les données proviennent d'une loi normale $\mathcal{N}(\mu, \sigma^2)$ de moyenne inconnue, mais de **variance connue**. Oui, ce cadre est assez restrictif, mais ce test a une vocation pédagogique, car on a besoin de rien connaître quasiment, à part les lois normales (pas besoin de connaître la loi de Student ici). Cette section est donc un préambule pour le test de Student, qui est similaire, sauf que l'on suppose que la variance est inconnue.

On souhaite tester si $\mu = \mu_0$ pour une certaine constante μ_0 .

$$\begin{aligned} H_0 : & \mu = \mu_0 \\ H_1 : & \mu \neq \mu_0 \end{aligned}$$

Exemple 302.

$$\begin{aligned} H_0 : & \mu = 42\text{cm} \text{ (les longueurs de sardines sont en moyenne de } 42\text{cm)} \\ H_1 : & \mu \neq 42\text{cm} \end{aligned}$$

On considère un échantillon (X_1, \dots, X_n) où $X_i \sim \mathcal{N}(\mu, \sigma^2)$. On pose la moyenne :

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

Afin d'avoir une loi normale centrée réduite sous l'hypothèse H_0 , on définit la statistique suivante :

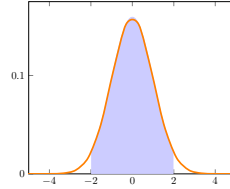
$$Z(X_1, \dots, X_n) := \sqrt{n} \frac{\bar{X}_n - \mu_0}{\sigma}.$$

Proposition 303. Sous l'hypothèse H_0 , $Z \sim \mathcal{N}(0, 1)$.

DÉMONSTRATION.

C'est la proposition 267. ■

On sait alors qu'on a 95% de chances d'être dans la zone bleue :



où les bords -1.959964 et 1.959964 sont respectivement les quantiles $q_{0.025} = q_{\alpha/2}$ et $q_{0.975} = q_{1-\alpha/2}$ pour $\alpha = 5\%$.

Proposition 304. *Sous l'hypothèse H_0 , $Z \sim \mathcal{N}(0, 1)$. Alors :*

- $\mathbb{P}(Z \leq q_{\alpha}) = \alpha$
- $\mathbb{P}(Z \geq q_{1-\alpha}) = \alpha$
- $\mathbb{P}(|Z| \geq q_{1-\alpha/2}) = \alpha$

```

fonction testZ( $x_1, \dots, x_n$ )
    calculer  $Z(x_1, \dots, x_n)$  en fonction des observations
    soit  $q_{1-\alpha/2}$  le quantile d'ordre  $1 - \alpha/2$  de  $\mathcal{N}(0, 1)$ 
    si  $Z \in [-q_{1-\alpha/2}, q_{1-\alpha/2}]$ 
        |   conserver  $H_0$ 
    sinon
        |   rejeter  $H_0$ 

```

Théorème 305. $\mathbb{P}(\text{rejeter } H_0 \mid H_0) = \alpha$.

Dans le théorème précédent et dans la suite, pour simplifier on utilise une notation conditionnelle. Mais ça n'est pas une. En écrivant ' $\mathbb{P}(\text{rejeter } H_0 \mid H_0) = \alpha$ ', on veut dire, 'si H_0 est supposée vraie, alors $\mathbb{P}(\text{rejeter } H_0) = \alpha$ '.

Proposition 306. *Si H_0 est fausse (et donc H_1 est vraie) alors, $\mathbb{P}(\text{rejeter } H_0 \mid H_1) = \mathbb{P}(|Z| \leq q_{1-\alpha/2})$ où Z suit une loi $\mathcal{N}(\mu - \mu_0, 1)$.*

La **p-value** est la probabilité d'observer une statistique de test qui est au moins aussi extrême que celle obtenue avec les données, en supposant H_0 . Ici, elle est définie comme suit.

Définition 307. *Soit x_1, \dots, x_n les données. On note z la valeur obtenu pour Z , c'est-à-dire $z = Z(x_1, \dots, x_n)$. La **p-value** est donnée par*

$$\mathbb{P}(|Z| \geq |z| \mid H_0).$$

14.6 Test de Student

On présente le même cadre que le test z , mais avec une variance inconnue. On sait que les données proviennent d'une loi normale $\mathcal{N}(\mu, \sigma^2)$, mais à la fois la moyenne μ et la variance σ^2 sont inconnues. On souhaite tester si $\mu = \mu_0$ pour une certaine constante μ_0 .

$$\begin{aligned} H_0 : & \mu = \mu_0 \\ H_1 : & \mu \neq \mu_0 \end{aligned}$$

On considère un échantillon (X_1, \dots, X_n) où $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

On pose la moyenne :

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

On pose aussi l'estimateur sans biais de la variance (cf. définition 229 page 75) :

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

On définit la statistique

$$Z = \sqrt{n} \frac{\bar{X}_n - \mu_0}{S_n^*}.$$

Proposition 308. Si H_0 est vraie, alors $Z \sim t(n-1)$ où $t(n-1)$ est la loi de Student à $n-1$ degré de liberté (cf. définition 270 page 91).

DÉMONSTRATION.

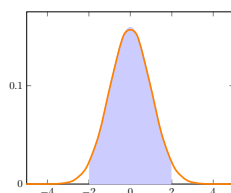
Même démonstration que la proposition proposition 271 page 91. ■

Proposition 309. Si H_0 est fausse, alors

Proposition 310. Soit $Z \sim t(n-1)$. Alors :

- $\mathbb{P}(Z \leq q_\alpha) = \alpha$
- $\mathbb{P}(Z \geq q_{1-\alpha}) = \alpha$
- $\mathbb{P}(|Z| \geq q_{1-\alpha/2}) = \alpha$

fonction test t de Student(x_1, \dots, x_n)
 calculer Z en fonction des observations
 soit $q_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi $t(n-1)$
 si $|Z| \leq q_{1-\alpha/2}$ conserver H_0 sinon rejeter H_0



14.7 Métaphore du procès

Il existe plusieurs tests statistiques. Ils varient selon la nature de l'hypothèse H_0 . On verra des exemples plus tard. Mais on a d'abord besoin de comprendre ce qu'est un test en général.

14.7.1 Le procès

Un test statistique peut être vu comme un juge dans le **procès** suivant. L'hypothèse nulle H_0 est l'accusée. Elle est accusée d'être fausse par la chercheuse. Le test doit juger si H_0 est innocente (i.e. toujours conservée comme vraie), ou alors coupable (i.e. rejetée comme étant vraie). La chercheuse vient avec toutes les données pour que le juge rejette H_0 . De son côté, H_0 a la présomption d'innocence. Il n'est pas souhaitable de mettre une hypothèse vraie en prison.

hypothèse nulle	=	accusée
données	=	témoin
juge	=	algorithme de test statistique

L'idée d'un test est de conserver/rejeter H_0 . Le test rejette H_0 si il y a suffisamment de données qui 'contredisent' H_0 . Attention : conserver H_0 ne signifie pas H_0 soit vraie. Cela signifie juste que l'on ne dispose pas de suffisamment d'informations pour contredire H_0 . C'est la **présomption d'innocence**.

Exemple 311. Avec deux lancers 🎲 🎲, on continue à conserver l'hypothèse H_0 que le dé est équilibré. Deux lancers ne suffisent pas à conclure.

Exemple 312. Avec 🎲🎲🎲🎲🎲🎲🎲🎲🎲🎲🎲🎲🎲🎲, on commence à être vraiment suspicieux.

Généralement, on conserve H_0 si on ne dispose pas de suffisamment de données pour affirmer le contraire.

Définition 313. On dit que le test a produit un **résultat statistiquement significatif** si l'hypothèse nulle a été rejeté à partir des données d'une expérience.

▲ La terminologie 'statistiquement significatif' est ancienne. Elle ne signifie pas que le résultat est important, mais que le résultat est 'signifié', juste que les données ont permis rejeter l'hypothèse nulle.

Définition 314. T s'appelle la **statistique de test** (ou parfois **variable de décision**).

Définition 315 (région de rejet). La région de rejet R est l'ensemble tel que si la statistique prend une valeur dans R , on décide de rejeter H_0 .

14.7.2 Erreurs de jugement

Le tableau suivant reporte les différents cas de figures. Le test peut tout à fait très bien fonctionner et on fait une bonne décision (représentée par les ✓ dans le tableau). Mais on peut malheureusement aussi se tromper : ce sont les erreurs de type 1 et celles de type 2.

	H_0 est vraie Amélie est innocente	H_1 est vraie Amélie est coupable
conserver H_0 Amélie sera en liberté	✓ innocente et en liberté	✗ erreur de type 2 coupable et en liberté
rejeter H_0 Amélie sera en prison	✗✗ erreur de type 1 innocente et en prison	✓ coupable et en prison

Remarque 316. L'erreur de type 1 s'appelle parfois aussi erreur de première espèce. C'est une erreur grave que de mettre en prison une personne innocente. L'erreur de type 2 s'appelle parfois aussi erreur de seconde espèce.

Désolé...

Type 1 et type 2 n'est pas vraiment une dénomination très informative. Mais c'est bien la terminologie utilisée en statistiques, et nous ne pouvons pas maintenant la changer. Mais on s'en souvient facilement. Type 1, c'est le plus dangereux : mettre une innocente en prison c'est horrible. Un voleur en liberté c'est moins grave (qui n'a jamais volé de dentifrice dans un supermarché et pourtant n'a pas eu de peine?).

14.7.3 Compromis entre erreurs de type 1 et 2

Définition 317 (risque de première espèce, risque de deuxième espèce).

Niveau de
signification
du test = Risque de 1ère espèce = α = $\mathbb{P}(\text{erreur de type 1})$ = $\mathbb{P}(\text{rejeter } H_0 \mid H_0 \text{ vraie})$

Risque de 2e espèce = β = $\mathbb{P}(\text{erreur de type 2})$ = $\mathbb{P}(\text{conserver } H_0 \mid H_1 \text{ vraie})$

⚠ Attention, pour plus de clarté, nous avons utilisé la notation d'une probabilité conditionnelle, e.g. $\mathbb{P}(\text{rejeter } H_0 \mid H_0 \text{ vraie})$. Mais ce n'est pas une probabilité conditionnelle, tout simplement car ' H_0 est vraie' n'est pas un événement ! H_0 est soit vraie, soit fausse. Par contre, dans les tests bayésiens, ' H_0 est vraie' est un événement et on pourra écrire cela comme cela.

Définition 318 (puissance du test).

$$\text{puissance} = 1 - \beta = \mathbb{P}(\text{rejeter } H_0 \mid H_1 \text{ vraie})$$

On indique les probabilités dans le tableau :

	H_0 est vraie Amélie est innocente	H_1 est vraie Amélie est coupable
conserver H_0 Amélie sera en liberté	✓ $(1 - \alpha)$ innocente et en liberté	✗ β coupable et en liberté (faux négatif)
rejeter H_0 Amélie sera en prison	✗✗ α innocente et en prison (faux positif)	✓ $(1 - \beta)$ coupable et en prison

14.7.4 Démarche commune

1. Fixer un seuil pour le niveau de signification, e.g. $\alpha \leq 5\%$
2. Maximiser la puissance $1 - \beta$ du test

⚠ Attention, la p -value n'est pas la probabilité que H_0 soit vraie. C'est une erreur courante chez les scientifiques. On veut tellement quantifier de combien H_0 est vraie... bref, on aimerait être bayésien (on l'est peut-être non ?). Mais les tests statistiques ici sont fréquentistes.

14.8 Test t de Student pour échantillons indépendants

Cadre : deux échantillons (X_1, \dots, X_n) et (Y_1, \dots, Y_m) avec $X_i \sim \mathcal{N}(\mu_1, \sigma)$ et $Y_j \sim \mathcal{N}(\mu_2, \sigma)$

▲ On suppose que les deux échantillons ont même variances σ . C'est stupide, on verra le test t de Welsh juste après.

$$\begin{aligned} H_0 : & \mu_1 = \mu_2 \\ H_1 : & \mu_1 \neq \mu_2 \end{aligned}$$

14.9 Test t de Welsh pour échantillons indépendants

Cadre : deux échantillons (X_1, \dots, X_n) et (Y_1, \dots, Y_m) avec $X_i \sim \mathcal{N}(\mu_1, \sigma_1)$ et $Y_j \sim \mathcal{N}(\mu_2, \sigma_2)$

$$\begin{aligned} H_0 : & \mu_1 = \mu_2 \\ H_1 : & \mu_1 \neq \mu_2 \end{aligned}$$

14.10 Test de Kolmogorov-Smirnov

Ce test est souvent utilisé pour tester qu'un générateur aléatoire de nombres vérifie une certaine distribution représentée par une fonction de répartition continue $F : \mathbb{R} \rightarrow [0, 1]$.

Exemple 319. Est-ce que les données suivantes proviennent d'une loi normale centrée et réduite ?

6.42, 5.23, -1.25, 0.12, -0.01, -1.02, 18.54, 0.06, -7.64, 2.85, -1.84, 0.74, -0.65, 0.24

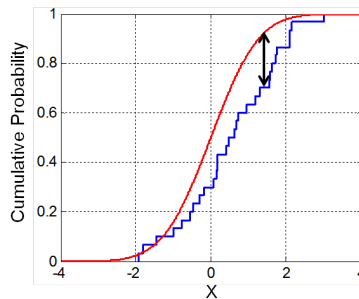
H_0 : le générateur/phénomène suit une loi de fonction de répartition $F : \mathbb{R} \rightarrow [0, 1]$

H_1 : le générateur suit une autre loi

— entrée : données $x_1, \dots, x_n \in \mathbb{R}$ générées par le générateur

— sortie : conserver H_0 , ou rejeter H_0 .

On construit une fonction de répartition empirique $F_{x_1, \dots, x_n}(x)$ à partir des données x_1, \dots, x_n en attribuant une probabilité de $\frac{1}{n}$ à chacune des données. Le test consiste ensuite à comparer la fonction de répartition théorique F (en rouge) avec la fonction de répartition empirique $F_{x_1, \dots, x_n}(x)$ (en bleu).



La fonction de répartition empirique (en bleu) F_{x_1, \dots, x_n} est définie comme suit. Comme $F_{x_1, \dots, x_n}(x)$ est censé être la probabilité d'avoir une valeur plus petit que x , on regarde la proportion de x_i plus petit que x , i.e. le nombre de x_i plus petit que x divisé par n .

Définition 320. Étant donné des données $x_1, \dots, x_n \in \mathbb{R}$, on définit la **fonction de répartition empirique** :

$$F_{x_1, \dots, x_n}(x) = \frac{\#\{i \in \{1, \dots, n\} \mid x_i \leq x\}}{n}.$$

La statistique utilisée est la distance de convergence uniforme entre F_{x_1, \dots, x_n} et F .

Définition 321. La statistique est :

$$T(x_1, \dots, x_n) = \sqrt{n} \|F_{x_1, \dots, x_n} - F\|_\infty$$

où $\|F_{x_1, \dots, x_n} - F\|_\infty = \sup_{x \in \mathbb{R}} |F_{x_1, \dots, x_n}(x) - F(x)|$.

Pont brownien


Il y a un rapport entre la différence entre F et F_{x_1, \dots, x_n} et la notion de pont brownien.

Définition 322. La **loi de Kolmogorov-Smirnov** μ_{KS} est donnée par la fonction de répartition

$$F_{KS}(x) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2 x^2}.$$

Théorème 323 (admis). Si X_1, \dots, X_n sont iid et de loi de fonction de répartition F continue, alors

$$T(X_1, \dots, X_n) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mu_{KS}.$$

 **Exemple 324.** Pour les valeurs citées plus haut, on calcule : $T(x_1, \dots, x_n) = 0.2521$.

Pour $n = 15$, on a environ que

$$\mathbb{P}[\sup_{x \in \mathbb{R}} |F_{X_1, \dots, X_n}(x) - F(x)| > 0.304] \approx 0.10.$$

On conserve donc l'hypothèse nulle.

14.11 Méthode de Neyman et Pearson (*)

On considère des hypothèses sous la forme

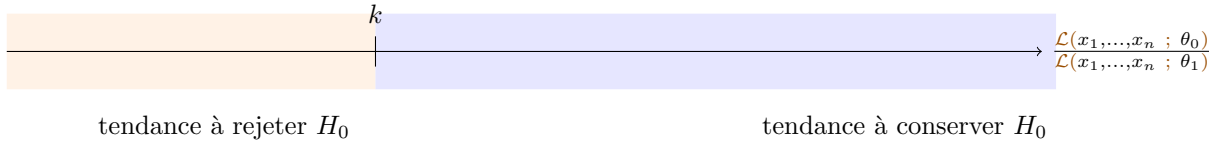
$$\begin{aligned} H_0 : & \theta = \theta_0 \\ H_1 : & \theta = \theta_1 \end{aligned}$$

On construit un test où la statistique utilisée est basé sur la vraisemblance. On rappelle que la vraisemblance mesure de combien des données sont vraisemblables par rapport à une distribution donnée (cf. Définition 233).

Les données recueillies sont x_1, \dots, x_n . On calcule $\mathcal{L}(x_1, \dots, x_n ; \theta_0)$ qui est la vraisemblance des données par rapport à la loi correspondante à \mathbb{P}_{θ_0} . On calcule aussi $\mathcal{L}(x_1, \dots, x_n ; \theta_1)$ qui est la vraisemblance des données par rapport à la loi correspondante à \mathbb{P}_{θ_1} . Puis on calcule le rapport

$$\frac{\mathcal{L}(x_1, \dots, x_n ; \theta_0)}{\mathcal{L}(x_1, \dots, x_n ; \theta_1)}$$

Plus ce rapport est petit, plus on a tendance à rejeter H_0 . A contrario, plus il est grand, et plus on a tendance à conserver H_0 . On montre cela sur l'axe suivant :



Reste maintenant à savoir où placer la limite k , qui détermine la région de rejet. Nous voulons que

$$\mathbb{P}(\text{rejeter } H_0 \mid H_0 \text{ vraie}) = \alpha.$$

Or on a

$$\begin{aligned} \mathbb{P}(\text{rejeter } H_0 \mid H_0 \text{ vraie}) &= \mathbb{P}\left(\frac{\mathcal{L}(x_1, \dots, x_n ; \theta_0)}{\mathcal{L}(x_1, \dots, x_n ; \theta_1)} \leq k \mid \theta = \theta_0\right) \\ &= \mathbb{P}\left(\frac{\mathcal{L}(X_1, \dots, X_n ; \theta_0)}{\mathcal{L}(X_1, \dots, X_n ; \theta_1)} \leq k \mid \theta = \theta_0\right). \end{aligned}$$

Test de puissance maximum(α, x_1, \dots, x_n)

1. Soit k tel que

$$\mathbb{P}\left(\frac{\mathcal{L}(X_1, \dots, X_n ; \theta_0)}{\mathcal{L}(X_1, \dots, X_n ; \theta_1)} \leq k \mid \theta = \theta_0\right) = \alpha$$

2. rejeter H_0 si

$$\frac{\mathcal{L}(x_1, \dots, x_n ; \theta_0)}{\mathcal{L}(x_1, \dots, x_n ; \theta_1)} \leq k$$

Lemme 325 (de Neyman et Pearson). Le test de puissance maximum garantit une puissance maximale sous la contrainte que le risque de première espèce $\leq \alpha$.

Chapitre 15

Tests du χ^2

15.1 Motivation

On s'intéresse à répondre à ces genres de problèmes.

Adéquation à une loi \mathcal{L} donnée

entrée : des données

sortie : les données suivent-elles la loi \mathcal{L} ? (H_0)

Test d'indépendance

entrée : des données

sortie : est-ce qu'on remarque une indépendance dans les données (H_0)?

Pour chacun de ces problèmes, nous allons mesurer la qualité des données (oui, elle a l'air de suivre la loi \mathcal{L} , oui, a priori, on remarque une indépendance, etc.). Pour cela, nous introduisons une statistique (i.e. variables aléatoires) qui ont la forme d'une distance au carrée :

$$D := (\text{perçu} - \text{idéal}_{H_0})^2 + (\text{perçu}' - \text{idéal}'_{H_0})^2 + \dots$$

où chaque terme $(\text{perçu} - \text{idéal}_{H_0})$ suit une loi normale centrée (ou alors est approximée par une loi normale centrée).

Dans un monde idéal $D = 0$. On espère avoir D petit. Si par contre, si D est trop grand, on remet en cause notre hypothèse H_0 .

Une telle statistique D suit la loi dite du χ^2 avec k degrés de liberté, où

$$k = \text{nombre de termes de la forme } (\text{perçu} - \text{idéal})^2 - \text{le nombre de relations qui les lient.}$$

Ainsi, on peut savoir quand D est trop grand : c'est quand

$$\underbrace{\mathbb{P}(\text{avoir ces données ou pire} \mid H_0)}_{\text{p-value}} \leq \alpha$$

où α est un seuil. On évalue cette probabilité à l'aide de la loi du χ^2 .

15.2 Test d'adéquation à une loi donnée

On sait que les données suivent une loi discrète \mathbf{p} . On pense très fortement que la loi suivie est donnée par \mathbf{p}^0 . C'est notre hypothèse H_0 .

$$\begin{aligned} H_0 : & \mathbf{p} = \mathbf{p}^0 \\ H_1 : & \mathbf{p} \neq \mathbf{p}^0 \end{aligned}$$

Adéquation à la loi \mathbf{p}^0

entrée : des observations

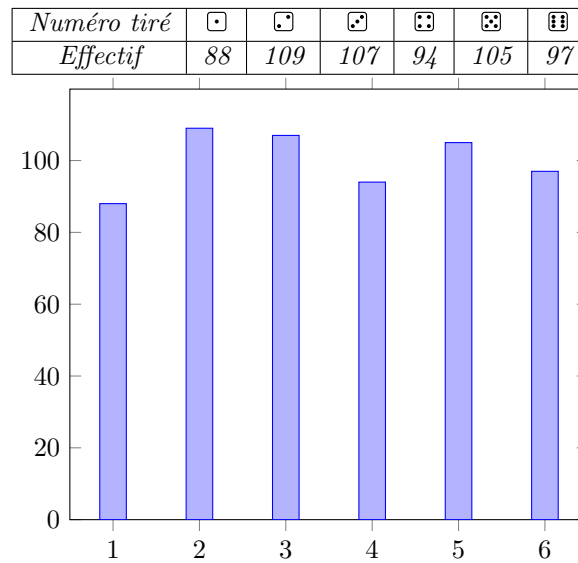
sortie : est-ce que les observations suivent la loi \mathbf{p} ?

Exemple 326 (dé non pipé?).

H_0 : dé non pipé, i.e. $\mathbf{p} = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$

H_1 : dé pipé $\mathbf{p} \neq (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$

Sur $n = 600$ lancers, on observe :



Est-ce que le dé est bien non pipé ?

15.2.1 Statistique utilisée

On reprend les notations de la section 8.3 :

- On considère X_1, \dots, X_n des variables aléatoires à valeurs dans $\{1, \dots, k\}$, iid selon la distribution \mathbf{p} .
- On note $N_i(n)$ la variable aléatoire qui vaut le nombre de fois où l'on obtient un élément dans la classe i sur les premiers n lancers aléatoires.

Exemple 327 (lancers d'un dé). On a $k = 6$.

X_t = la valeur du dé du t -ème lancer

$N_i(n)$ = nombre de i obtenus sur les n lancers

La définition qui suit donne la statistique utilisée pour faire le test. C'est cette statistique dont on va étudier la loi. Elle est définie comme une mesure d'erreurs par rapport à l'idéal. La quantité $N_i(n)$ est le nombre "réel" de fois que l'on tombe dans la classe i , alors que np_i est le nombre "théorique" où l'on tombe dans la classe i .

Définition 328. La *statistique utilisée pour faire le test* est :

$$\begin{aligned}
 D_n &:= \sum_{i=1}^k \frac{(nb \text{ d'éléments dans la classe } i - nb \text{ d'éléments théorique dans la classe } i \text{ sous } H_0)^2}{nb \text{ d'éléments théorique dans la classe } i \text{ sous } H_0} \\
 &= \sum_{i=1}^k \frac{(N_i(n) - np_i)^2}{np_i} \\
 &= n \sum_{i=1}^k \frac{(\frac{N_i(n)}{n} - p_i)^2}{p_i}
 \end{aligned}$$

Exemple 329 (n lancers d'un dé). Pour n lancers de dé, on obtient :

$$\begin{aligned}
 D_n = n \Big[& \frac{(\frac{\# \text{ de } \square \text{ obtenus}}{n} - 1/6)^2}{1/6} + \frac{(\frac{\# \text{ de } \square \text{ obtenus}}{n} - 1/6)^2}{1/6} + \frac{(\frac{\# \text{ de } \square \text{ obtenus}}{n} - 1/6)^2}{1/6} + \\
 & \frac{(\frac{\# \text{ de } \square \text{ obtenus}}{n} - 1/6)^2}{1/6} + \frac{(\frac{\# \text{ de } \square \text{ obtenus}}{n} - 1/6)^2}{1/6} + \frac{(\frac{\# \text{ de } \square \text{ obtenus}}{n} - 1/6)^2}{1/6} \Big]
 \end{aligned}$$

On rappelle que la matrice de covariance de $N(n)$ est $n(\Delta_{\mathbf{p}} - \mathbf{p}\mathbf{p}^t)$ (corollaire 186). On rappelle que l'application du TCL (cf. corollaire 207) donne :

$$\sqrt{n}\left(\frac{N(n)}{n} - \mathbf{p}\right) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, \Delta_{\mathbf{p}} - \mathbf{p}\mathbf{p}^t).$$

15.2.2 Comportement de la statistique quand H_0 fausse

Théorème 330 (quand H_0 est fausse). *Supposons que H_0 fausse, i.e. $(X_n)_{n \in \mathbb{N}}$ une suite de v.a. iid de loi \vec{q} avec \vec{q} un vecteur proba différent de \mathbf{p} . Alors :*

$$D_n \xrightarrow[n \rightarrow +\infty]{p.s.} +\infty.$$

DÉMONSTRATION.

En appliquant la loi des grands nombres, on obtient

$$\frac{N_i(n)}{n} \xrightarrow[n \rightarrow +\infty]{p.s.} q_i.$$

Ainsi :

$$\sum_{i=1}^k \frac{\left(\frac{N_i(n)}{n} - p_i\right)^2}{p_i} \xrightarrow[n \rightarrow +\infty]{p.s.} \sum_{i=1}^k \frac{(q_i - p_i)^2}{p_i}.$$

Comme $\vec{q} \neq \mathbf{p}$, on a :

$$\sum_{i=1}^k \frac{(q_i - p_i)^2}{p_i} > 0$$

Comme D_n vaut n fois la quantité ci-dessus, i.e. $D_n = n \sum_{i=1}^k \frac{\left(\frac{N_i(n)}{n} - p_i\right)^2}{p_i}$, on a $D_n \xrightarrow[n \rightarrow +\infty]{p.s.} +\infty$.

■

Remarquez que si l'hypothèse H_0 était vraie ($\mathbf{p} = \vec{q}$), on aurait

$$\sum_{i=1}^k \frac{\left(\frac{N_i(n)}{n} - p_i\right)^2}{p_i} \xrightarrow[n \rightarrow +\infty]{p.s.} 0$$

et D_n est un zoom sur cette convergence vers 0. Bref... on va utiliser le TCL.

15.2.3 Comportement de la statistique quand H_0 vraie

Théorème 331 (quand H_0 est vraie). *Soit $(X_n)_{n \in \mathbb{N}}$ une suite de v.a. iid de loi \mathbf{p} . Alors*



$$D_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi^2(k-1).$$

DÉMONSTRATION.

On pose la fonction

$$f : \vec{z} \mapsto \frac{z_1^2}{p_1} + \dots + \frac{z_k^2}{p_k}.$$

Voici le squelette de la démonstration.

Theorème centrale limite vectoriel
(Cor. 207)

$\sqrt{n}\left(\frac{N(n)}{n} - \mathbf{p}\right) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, \Delta_{\mathbf{p}} - \mathbf{p}\mathbf{p}^t)$	Définition de D_n Définition de f	lemme 332
$f\left(\sqrt{n}\left(\frac{N(n)}{n} - \mathbf{p}\right)\right) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} f(\mathcal{N}(0, \Delta_{\mathbf{p}} - \mathbf{p}\mathbf{p}^t))$	Prop. 157	$f(\sqrt{n}\left(\frac{N(n)}{n} - \mathbf{p}\right)) = D_n$
$D_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi^2(k-1)$		

Lemme 332. *Si $Z \sim \mathcal{N}(0, \Delta_{\mathbf{p}} - \mathbf{p}\mathbf{p}^t)$ implique $\frac{Z_1^2}{p_1} + \dots + \frac{Z_k^2}{p_k} \sim \chi^2(k-1)$.*



DÉMONSTRATION.

On note $U_i = \frac{Z_i}{\sqrt{p_i}}$. On pose $U = (U_1, \dots, U_k)$. Par la proposition 114, on a $\text{cov}(U_i U_j) = \frac{1}{\sqrt{p_i} \sqrt{p_j}} \text{cov}(Z_i Z_j)$. Ainsi :

$$U \sim \mathcal{N}(0, I - \sqrt{p} \sqrt{p}^t).$$

La matrice $I - \sqrt{p} \sqrt{p}^t$ est la matrice de projection orthogonal sur l'hyperplan orthogonal au vecteur \sqrt{p} . En effet, c'est bien la matrice de l'application linéaire projection orthogonal sur hyperplan orthogonal au vecteur \sqrt{p} définie par :

$$\text{proj}(x) = x - \langle x, \sqrt{p} \rangle \sqrt{p}.$$

Il existe donc une matrice orthogonale O tel que

$$O(I - \sqrt{p} \sqrt{p}^t)O^t = \Delta_{1, \dots, 1, 0}$$

où $\Delta_{1, \dots, 1, 0}$ est la matrice diagonale de taille $k \times k$, avec que des 1 sur la diagonale sauf tout en bas à droite où il y a un 0.

On a :

$$OU \sim \mathcal{N}(0, \Delta_{1, \dots, 1, 0}).$$

En effet, la matrice de covariance de OU est

$$\mathbb{E}((OU)(OU)^t) = \mathbb{E}(OUU^tO^t) = O\mathbb{E}(UU^t)O^t = O\text{cov}(U)O^t = \Delta_{1, \dots, 1, 0}.$$

Dit autrement, les $k-1$ premières coordonnées de OU suivent des lois normales centrées réduites et indépendantes, alors que la dernière coordonnées de OU est nulle.

On a

$$\frac{Z_1^2}{p_1} + \dots + \frac{Z_k^2}{p_k} = \|U\|^2 = \|OU\|^2.$$

Dit autrement, $\frac{Z_1^2}{p_1} + \dots + \frac{Z_k^2}{p_k}$ peut s'écrire comme une somme de $k-1$ variables aléatoires indépendantes, toutes suivant une loi normale réduite. Par définition de $\chi^2(k-1)$, on a :

$$\frac{Z_1^2}{p_1} + \dots + \frac{Z_k^2}{p_k} \sim \chi^2(k-1).$$

■ ■

15.2.4 Algorithme

Le test du χ^2 repose sur les deux théorèmes précédents. Autrement dit, il s'appuie sur des résultats asymptotiques, alors que l'on effectue qu'un nombre fini d'observations. En pratique, la recette magique est que le test est valide quand $np_i \geq 5$ pour tout $i \in 1, \dots, k$. Si tel n'est pas le cas, on regroupe des classes : voiture + camion = véhicule.

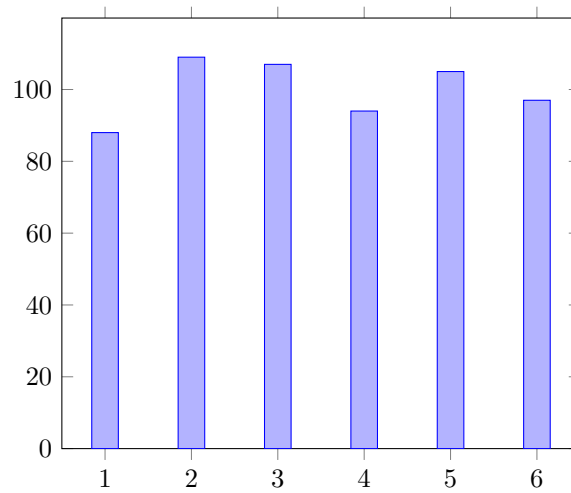
calculer D_n en fonction des observations
soit $q_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi $\chi^2(k-1)$
si $D_n \leq q_{1-\alpha}$ conserver H_0 **sinon** rejeter H_0

Exemple 333 (dé pipé ou non?).

H_0 : dé non pipé, i.e. $\mathbf{p} = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$
 H_1 : dé pipé $\mathbf{p} \neq (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$

Sur n lancers, on observe :

Numéro tiré	□	□	□	□	□	□
Effectif	88	109	107	94	105	97



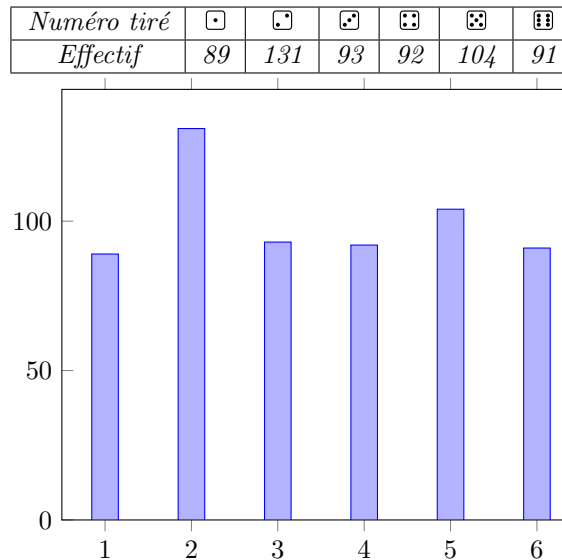
$$D_n = \frac{(88-100)^2}{100} + \frac{(109-100)^2}{100} + \frac{(107-100)^2}{100} + \frac{(94-100)^2}{100} + \frac{(105-100)^2}{100} + \frac{(97-100)^2}{100} = 3.44$$

Voici le quantile d'ordre 0.95 de la loi du $\chi^2(5)$: $q_{0.95} = 11.07$.

Comme $D_n < q_{0.95}$, on conserve H_0 .

Exemple 334 (dé pipé ou non ? Autre exemple).

Maintenant sur $n = 600$ lancers on observe :



$$D_n = \frac{(89-100)^2}{100} + \frac{(131-100)^2}{100} + \frac{(93-100)^2}{100} + \frac{(92-100)^2}{100} + \frac{(104-100)^2}{100} + \frac{(91-100)^2}{100} = 12.92$$

Ici $D_n \geq q_{0.95}$, on rejette H_0 .

Remarque 335. Le test du χ^2 fait comme si on avait atteint les limites en loi et ps. Les livres préconisent $np_i \geq 5$ pour que le test soit valide. Sinon, on regroupe des classes.

15.3 Test d'adéquation à une famille de lois

On se fixe une famille de paramètres $\Theta \subseteq \mathbb{R}^d$. On fixe une famille de lois $(\mathbf{p}(\theta))_{\theta \in \Theta}$.

Adéquation à une famille de lois

entrée : des données

sortie : est-ce que les données respectent une loi $\mathbf{p}(\theta)$ pour un certain θ ?

H_0 : la probabilité réelle \mathbf{p} est $\mathbf{p}(\theta)$ pour un certain θ dans Θ

H_1 : la probabilité réelle \mathbb{P} est différente de $\mathbf{p}(\theta)$ pour tout $\theta \in \Theta$

15.3.1 Statistique

La statistique utilisée est quasiment la même que pour l'adéquation à une loi donnée, sauf que, au lieu d'utiliser \mathbf{p} , on utilise $\mathbf{p}(\theta)$ pour la valeur du paramètre θ la plus vraisemblable, autrement dit... pour l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ (voir Définition 235).

Définition 336. Soit $N_i(n)$ la variable aléatoire qui vaut le nombre de fois où l'on obtient un élément dans la classe i sur n lancers aléatoires. On pose :

$$D_n := \sum_{i=1}^k \frac{(N_i(n) - np_i(\hat{\theta}_n))^2}{np_i(\hat{\theta}_n)}$$

où $\hat{\theta}_n$ est l'estimateur du maximum de vraisemblance de θ .

Théorème 337 (admis). Soit $(X_n)_{n \in \mathbb{N}}$ une suite de va, iid de loi $\mathbf{p}(\theta)$ pour un certain $\theta \in \Theta$ (inconnu). Alors

$$D_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi^2(k - d - 1)$$

où k est le nombre de classes, d est le nombre de paramètres.

Théorème 338.

Si H_0 vraie, alors $D_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi^2(k - d - 1)$

Sinon $D_n \xrightarrow[n \rightarrow +\infty]{p.s.} +\infty$

15.3.2 Algorithme

calculer D_n en fonction des observations
 soit $q_{1-\alpha}$ le quantile d'ordre $1 - \alpha$ de la loi $\chi^2(k - d - 1)$
si $D_n \leq q_{1-\alpha}$ conserver H_0 **sinon** rejeter H_0

Exercice 15.1. Pour 10000 fratries de quatre enfants (exactement), on a relevé le nombre de garçons :

nombre de garçons	0	1	2	3	4
effectifs	572	2329	3758	2632	709

On modélise les naissances successives de la façon suivante.

- les naissances sont indépendantes ;
- à chaque naissance, la livraison est un garçon ou une fille avec probabilités respectives θ et $1 - \theta$.

1. Dans ce modèle, quelle est la loi p du nombre de garçons dans une fratrie de quatre enfants ?
2. Tester l'hypothèse $H_0 : p = \mathcal{B}(4, 1/2)$ contre $H_1 : p \neq \mathcal{B}(4, 1/2)$ au niveau 0,05.
3. Tester l'hypothèse $H_0 : p \in \{\mathcal{B}(4, \theta), \theta \in]0, 1[\}$ contre $H_1 : p \notin \{\mathcal{B}(4, \theta), \theta \in]0, 1[\}$.
4. Conclusion ?

Exercice 15.2. On étudie le nombre de connexion à Google pendant la durée de temps unitaire d'une seconde. On fait 200 mesures.

nombre de connexion par seconde	0	1	2	3	4	5	6	7	8	9	10	11
effectif empirique	6	15	40	42	37	30	10	9	5	3	2	1

Soit X la v.a. à valeurs dans \mathbb{N} comptant le nombre de connexions par seconde. Peut-elle être considérée comme une loi de Poisson au niveau 5% ?

15.4 Test d'indépendance

Soit $(Y_n, Z_n)_{n \in \mathbb{N}}$ v.a. iid. avec pour tout $n \in \mathbb{N}$, $Y_n \dots : \Omega \rightarrow \{1, \dots, r\}$ et $Z_n \dots : \Omega \rightarrow \{1, \dots, s\}$.

H_0 : Y_1 et Z_1 sont indépendantes

H_1 : Y_1 et Z_1 ne sont pas indépendantes

‘Pour toute classe $i \in \{1, \dots, r\}$, pour toute classe $j \in \{1, \dots, s\}$, on note $N_{i,j}$ le nombre de fois où on a eu (i, j) , $N_{i,\cdot}$ le nombre de fois où on a eu i , et $N_{\cdot,j}$ le nombre de fois où on a eu j . Formellement :

$$N_{i,j} = \#\{\ell \in \{1, \dots, n\} \mid (Y_\ell, Z_\ell) = (i, j)\}$$

$$N_{i,\cdot} = \sum_{j=1}^s N_{i,j}$$

$$N_{\cdot,j} = \sum_{i=1}^r N_{i,j}$$

Exemple 339 (efficacité de deux médicaments semblables mais de prix différents).

	Médicament cher	Médicament bon marché	
Guérisons	$N_{gc} = 156$	$N_{gb} = 44$	$N_{g\cdot} = 200$
Non-guérisons	$N_{nc} = 44$	$N_{nb} = 6$	$N_{n\cdot} = 50$
	$N_{\cdot c} = 200$	$N_{\cdot b} = 50$	$n = 250$

Y-a-t-il indépendance entre guérison et médicament cher ou pas ?

15.4.1 Statistique utilisée

Définition 340.

$$D_n := \sum_{i=1}^r \sum_{j=1}^s \frac{\left(N_{i,j} - \frac{N_{i,\cdot} N_{\cdot,j}}{n}\right)^2}{\frac{N_{i,\cdot} N_{\cdot,j}}{n}} = n \left(\sum_{i=1}^r \sum_{j=1}^s \frac{N_{ij}^2}{N_{i,\cdot} N_{\cdot,j}} - 1 \right)$$

Exemple 341. $D_n = 250 \left(\frac{156^2}{200 \times 200} + \frac{44^2}{50 \times 200} + \frac{44^2}{50 \times 200} + \frac{6^2}{50 \times 50} - 1 \right) = 2.5$

Théorème 342. Si H_0 vraie, alors $D_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi^2((r-1) \times (s-1))$ Sinon $D_n \xrightarrow[n \rightarrow +\infty]{p.s.} +\infty$

15.4.2 Algorithme

calculer D_n en fonction des observations
 soit $q_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi $\chi^2((r-1) \times (s-1))$
 si $D_n \leq q_{1-\alpha}$ conserver H_0 sinon rejeter H_0

Exemple 343. Le quantile de $\chi^2(1)$ d'ordre 0.95 vaut $q_{0.95} = 3.84$.

Comme $2.5 < 3.84$, on conserve H_0 .

Chapitre 16

Régression linéaire

16.1 Idée informelle

Régression

entrée : des features x_1, \dots, x_n , des étiquettes y_1, \dots, y_n

sortie : une fonction f qui colle au plus près des données, i.e. tel que $y_i \approx f(x_i)$ pour tout i

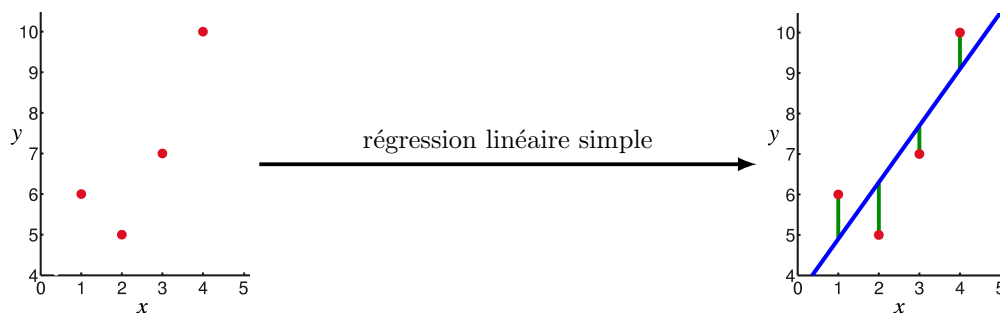
Régression linéaire

entrée : des features $x_1, \dots, x_n \in \mathbb{R}^p$, des étiquettes $y_1, \dots, y_n \in \mathbb{R}$

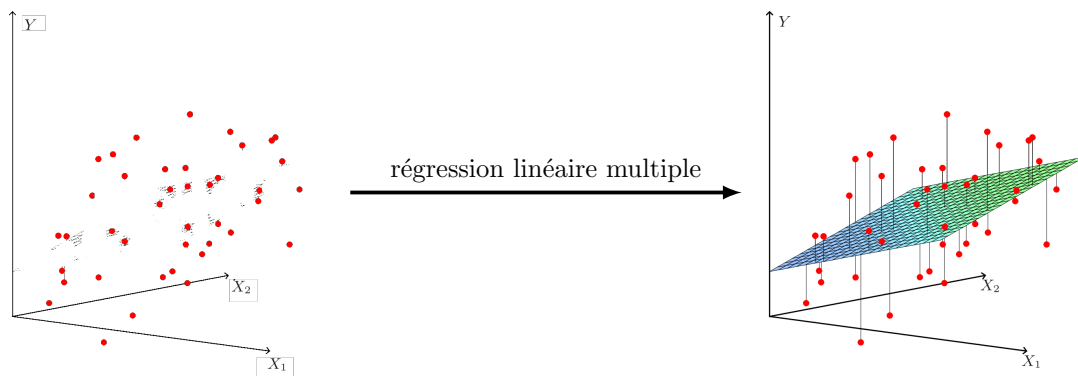
sortie : la droite/plan/hyper plan qui colle au plus près des données

Dit autrement, on cherche à expliquer y comme une combinaison linéaire des coordonnées de x .

Le cas le plus simple est le cas 2D : celui de rechercher une droite qui colle au plus près de points dans le plan. On parle de **régression linéaire simple**.



Les cas 3D, 4D, etc. i.e. en dimension quelconque s'appelle **régression linéaire multiple**. Par exemple, le cas 3D consiste à avoir des données dans l'espace. Chaque données est un point $((x_{i1}, x_{i2}), y_i) \in \mathbb{R}^2 \times \mathbb{R}$. On cherche alors le plan qui colle au plus près des points.



Le cas général est en dimension $p + 1$. Les points sont $((x_{i1}, \dots, x_{ip}), y_i) \in \mathbb{R}^p \times \mathbb{R}$.

16.2 Exemples d'applications



Expliquer $y = \text{prix d'une maison}$ en fonction de $x = \begin{pmatrix} \text{surface habitable} \\ \text{nombre de pièces} \\ \text{surface jardin} \\ \text{taux de criminalité du quartier} \\ \text{distance à l'école la plus proche} \end{pmatrix}$

Expliquer $y = \begin{matrix} \text{niveau de la protéine antigène} \\ \text{prostatique spécifique} \end{matrix}$ en fonction de $x = \begin{pmatrix} \text{âge} \\ \text{log poids de la prostate} \\ \text{log volume de la tumeur cancéreuse} \\ \vdots \end{pmatrix}$

Régression polynomiale

Régression polynomiale

entrée : des données $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$, un entier d

sortie : la surface polynomiale de degré au plus d qui colle au plus près des données

On peut faire une réduction algorithmique de la régression polynomiale vers la régression linéaire en ajoutant les monômes comme features :

$$x_1, \dots, x_n, x_1^2, \dots, x_n^2, x_1 x_2, \dots, x_1^3, \dots$$

16.3 Définition de la régression linéaire

On donne ici la définition générale de la régression linéaire, i.e. de la régression linéaire simple ($p = 1$) ou multiple ($p \geq 2$).

Définition 344 (Problème de la régression linéaire).

Régression linéaire

entrée : des données $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$;

sortie : $\theta = (\theta_0, \theta_1, \dots, \theta_p) \in \mathbb{R}^{p+1}$ qui minimise

$$\sum_{i=1}^n [y_i - (\theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \dots + \theta_p x_{ip})]^2.$$

Définition 345 (somme des moindres carrés). La quantité à minimiser s'appelle *somme des moindres carrés*, *sum of squared residuals (SSR)*, *error sum of squares (ESS)* ou *residual sum of squares (RSS)*. Autrement dit on pose :

$$RSS = \sum_{i=1}^n [y_i - (\theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \dots + \theta_p x_{ip})]^2.$$

Définition 346 (méthode des moindres carrés). On appelle *méthode des moindres carrés* le fait de minimiser RSS .

16.4 Régression linéaire simple

Commençons d'abord par étudier le cas où $p = 1$, i.e. le cas où chaque x_i est un nombre. Ce cas est plus simple à traiter mathématiquement : pas de notations matricielles.

Définition 347 (Problème de la régression linéaire simple). **Régression linéaire simple**

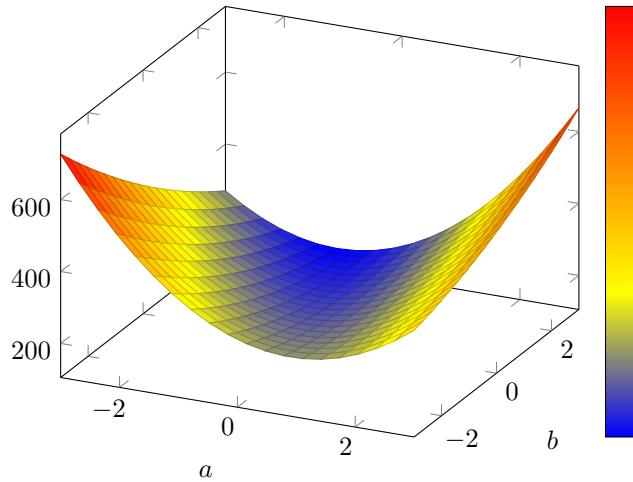
entrée : des données $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R} \times \mathbb{R}$;

sortie : $\theta = (a, b) \in \mathbb{R}^2$ qui minimise

$$\sum_{i=1}^n [y_i - (ax_i + b)]^2.$$

Exemple 348. Sur les données $(1, 6), (2, 5), (3, 7), (4, 10)$, la somme des moindres carrés est :

$$RSS = (6 - (a + b))^2 + (5 - (2a + b))^2 + (7 - (3a + b))^2 + (10 - (4a + b))^2.$$



Théorème 349. S'il existe i, j avec $x_i \neq x_j$, alors voici a et b qui minimise RSS :

$$a = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}$$

$$b = \bar{y}_n - a \times \bar{x}_n$$



où $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$ et $\bar{y}_n = \frac{1}{n} \sum_{i=1}^n y_i$.

DÉMONSTRATION.

Au minimum, le vecteur gradient de RSS est nul. Voici les coordonnées du vecteur gradient en tout point $(a, b) \in \mathbb{R}^2$:

$$\frac{\partial RSS}{\partial a} = -2 \sum_{i=1}^n x_i (y_i - ax_i - b)$$

$$\frac{\partial RSS}{\partial b} = -2 \sum_{i=1}^n (y_i - ax_i - b)$$

$\frac{dRSS}{db} = 0$ implique que $b = \bar{y}_n - a\bar{x}_n$. C'est déjà ça de trouvé. Maintenant, on va calculer a .

$\frac{dRSS}{da} = 0$ et en remplaçant b par $\bar{y}_n - a\bar{x}_n$, on obtient

$$\sum_i x_i y_i - \bar{y}_n \sum_i x_i = a \left(\sum_i x_i^2 - \bar{x}_n \sum_i x_i \right)$$

$$\sum_i x_i y_i - n \bar{y}_n \bar{x}_n = a \left(\sum_i x_i^2 - n \bar{x}_n^2 \right)$$

Or on a :

$$\begin{aligned}
- \sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n) &= \sum_{i=1}^n x_i y_i + n \bar{x}_n \bar{y}_n - \sum_{i=1}^n y_i \bar{x}_n - \sum_{i=1}^n x_i \bar{y}_n = \sum_{i=1}^n x_i y_i - n \bar{x}_n \bar{y}_n \\
- \sum_{i=1}^n (x_i - \bar{x}_n)^2 &= \sum_{i=1}^n x_i^2 + n \bar{x}_n^2 - 2 \sum_{i=1}^n x_i \bar{x}_n = \sum_{i=1}^n x_i^2 - n \bar{x}_n^2
\end{aligned}$$

On obtient donc l'expression suivante pour a :

$$a := \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}$$

■

16.5 Formulation matricielle

Revenons maintenant à la régression linéaire générale. Utilisons une formulation matricielle afin d'alléger les notations et prendre du recul. On remarque que le coefficient θ_0 est tout seul et n'est pas multiplié par une coordonnée des données. Pour ne pas s'embêter, on fait $\theta_0 = \theta_0 \times 1$, puis que ce 1 fait partie des données.

Définition 350 (Matrice des données). *La matrice des données est :*

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}.$$

Ainsi, chaque ligne correspond à une donnée différente (avec le 1 en plus en première colonne). Chaque colonne correspond à une feature (sauf la première colonne qui est la colonne des 1). Le problème de la régression linéaire se reformule alors comme cela.

Exemple 351 (régression linéaire simple). *Pour les données $(1, 6), (2, 5), (3, 7), (4, 10)$, la matrice des données est :*

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}.$$

Exemple 352 (régression linéaire en 3D). *Pour les données $((1, 4), 6), ((3, 5), 5), ((4, 8), 7), ((5, 11), 10)$, la matrice des données est :*

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 4 \\ 1 & 3 & 5 \\ 1 & 4 & 8 \\ 1 & 5 & 11 \end{pmatrix}.$$

Définition 353 (Problème de la régression linéaire).

Régression linéaire

entrée : la matrice $X \in \mathbb{R}^{n \times (p+1)}$ des données, et un vecteur colonne $y \in \mathbb{R}^n$;
sortie : trouver un vecteur colonne $\theta \in \mathbb{R}^{p+1}$ qui minimise

$$RSS = (y - \mathbf{X}\theta)^\top (y - \mathbf{X}\theta).$$

Dans la définition, n = nombre de points et p = dimension = nombre de descripteurs.

16.6 Solution

Théorème 354. *Si $\text{rang}(\mathbf{X}) = p + 1$, alors $\hat{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top y$ minimise le RSS.*

DÉMONSTRATION.

RSS est une forme quadratique convexe en θ . On la minimise en annulant son gradient, qui est :

$$\nabla_\theta RSS = -2\mathbf{X}^\top (y - \mathbf{X}\theta).$$

Comme $\text{rang}(\mathbf{X}) = p + 1$, $\mathbf{X}^\top \mathbf{X}$ est une matrice de taille $p + 1 \times p + 1$ inversible¹. On voit que θ^* annule $\hat{\theta}_{RSS}$. En effet :

$$\mathbf{X}^\top (y - \mathbf{X}\theta^*) = \mathbf{X}^\top y - \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top y = \mathbf{X}^\top y - \mathbf{X}^\top y = 0.$$

■

Remarque 355. Il se peut que $\text{rang}(\mathbf{X}) < p + 1$, et donc $\mathbf{X}^\top \mathbf{X}$ non inversible. Par exemple

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}$$

i.e l'unique descripteur de la régression linéaire simple vaut 0 pour toutes les points ($x_1 = x_2 = 0$). Ou alors

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \end{pmatrix}$$

est de rang $2 < 2 + 1$.

Dans ce cas, au lieu de $\mathbf{X}^\top \mathbf{X}$, on peut utiliser un pseudo-inverse, comme celui de Moore-Penrose. Aussi, la plupart des logiciels supprime alors les features redondantes.

- si p petit, on utilise la formule $\hat{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top y$
- si p grand, on utilise la méthode de gradient.

16.7 Théorème de Gauss-Markov

Dans cette section, nous allons faire le lien avec la théorie des estimateurs. Plus précisément, nous allons montrer que les moindres carrés donnent le ‘meilleur’ estimateur linéaire.

Définition 356 (BLUE). Soit (Y^1, \dots, Y^n) un échantillon aléatoire. Un meilleur estimateur linéaire non biaisé $\hat{\theta}$ de θ vérifie :

- $\hat{\theta}$ est linéaire, i.e. il s'écrit $\hat{\theta} = A \begin{pmatrix} Y^1 \\ \vdots \\ Y^n \end{pmatrix}$ avec $A \in \mathbb{R}^{n \times q}$;
- $\hat{\theta}$ est non biaisé, i.e. $\mathbb{E}(\hat{\theta}) = \theta$;
- la variance de tout estimateur linéaire non biaisé de θ est \geq à $\mathbb{V}(\hat{\theta})$.

≡ BLUE pour Best Linear Unbiased Estimator

Définition 357 (modèle linéaire). Le modèle linéaire est :

$$Y_i = \langle \theta, x_i \rangle + \epsilon_i$$

où les ϵ_i sont iid de même loi centrée et de variance σ^2 (connue ou non).

Théorème 358. Supposons que la matrice des données $\mathbf{X}^\top \mathbf{X}$ soit inversible. L'estimateur par la méthode des moindres carrés, i.e.

$$\hat{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \begin{pmatrix} Y^1 \\ \vdots \\ Y^n \end{pmatrix}$$

est l'unique BLUE de θ .

1. Pour le voir, primo on remarque que $\mathbf{X}^\top \mathbf{X}$ est symétrique. Comme $\text{rang}(\mathbf{X}) = p + 1$, le vecteur $\mathbf{X}v$ est non nul si v est non nul. Ainsi, $\mathbf{X}^\top \mathbf{X}$ est définie positive, et donc inversible.

16.8 Lien avec le maximum de vraisemblance

Dans cette section, on prend un modèle linéaire où tout est aléatoire et où le terme d'erreur soit une loi normale centrée.

Définition 359 (modèle de régression linéaire avec erreur gaussienne). *On considère un vecteur aléatoire $X : \Omega \rightarrow \mathbb{R}^p$, une variable aléatoire $Y : \Omega \rightarrow \mathbb{R}$, et une variable aléatoire $\epsilon : \Omega \rightarrow \mathbb{R}$ tels que*

$$Y = \langle \theta, X \rangle + \epsilon.$$

où $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

 **Théorème 360.** *Maximiser la vraisemblance des données $(x_i, y_i)_{i=1..n}$ revient à minimiser les moindres carrés.*

DÉMONSTRATION.

On note p la densité de (X, Y) . Les données sont $data = (x_i, y_i)_{i=1..n}$.

La densité s'écrit

$$\begin{aligned} p(x_i, y_i) &= p_{Y|X=x_i}(y_i) \times p_X(x_i) \\ &= p_{\mathcal{N}(0, \sigma^2)}(y_i - \langle \theta, x_i \rangle) \times p_X(x_i) \end{aligned}$$

Ainsi, la vraisemblance vaut :

$$\begin{aligned} \mathcal{L}(data ; \theta) &= \prod_{i=1}^n p(x_i, y_i) \\ &= \prod_{i=1}^n p_{\mathcal{N}(0, \sigma^2)}(y_i - \langle \theta, x_i \rangle) \times p_X(x_i) \\ &\propto \prod_{i=1}^n p_{\mathcal{N}(0, \sigma^2)}(y_i - \langle \theta, x_i \rangle) \\ &\propto \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y_i - \langle \theta, x_i \rangle)^2}{2\sigma^2}} \\ &\propto \prod_{i=1}^n e^{-\frac{(y_i - \langle \theta, x_i \rangle)^2}{2\sigma^2}} \\ &\propto e^{-\sum_{i=1}^n \frac{(y_i - \langle \theta, x_i \rangle)^2}{2\sigma^2}} \end{aligned}$$

où \propto signifie 'proportionnel à'. L'exposant est l'opposé de RSS. ■

16.9 Qualité d'une régression

On rappelle que nous avons calculer les coefficients $\hat{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top y$. Ainsi, les valeurs prédites pour y sont données par :

$$\hat{y} = \mathbf{X} \hat{\theta}.$$

En effet, c'est l'application linéaire (ou affine, mais comme la première colonne ne contient que des 1 c'est linéaire) de coefficient $\hat{\theta}$, à laquelle on applique chaque ligne de la matrice des données qui correspond à chaque x_i , $i = 1..n$.

Ainsi, si on pose $H = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$, on a

$$\hat{y} = Hy.$$

C'est pourquoi la matrice H s'appelle la matrice **hat** (chapeau), car elle ajoute le chapeau sur les valeurs prise par la variable expliquée (y devient \hat{y}).

Définition 361. *Sous le modèle linéaire (pas forcément gaussien, mais espérance nulle, et variance σ^2), alors voici un estimateur de σ^2 :*

$$\hat{\sigma}^2 := \frac{1}{n-p-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

Le dénominateur est $n-p-1$ et non n pour que l'estimateur soit sans biais.

Proposition 362. *L'estimateur $\hat{\sigma}^2$ de σ^2 est sans biais.*

Proposition 363. Soit θ les vraies valeurs du modèle linéaire et $\hat{\theta}$ l'estimateur de ces paramètres via les moindres carrés.

$$\hat{\theta} \sim \mathcal{N}(\theta, (\mathbf{X}^\top \mathbf{X})^{-1} \sigma^2)$$

Proposition 364.

$$(n - p - 1) \hat{\sigma}^2 \sim \sigma^2 \chi^2(n - p - 1)$$

.

Proposition 365. $\hat{\theta}$ et $\hat{\sigma}^2$ sont indépendants.

16.9.1 Tester qu'un coefficient est nul

Le fait qu'un coefficient θ_j est nul signifie que la j -ème feature n'explique rien du tout.

$$\begin{aligned} H_0 : & \theta_j = 0 \\ H_1 : & \theta_j \neq 0 \end{aligned}$$

Définition 366 (statistique de test).

$$z_j = \frac{\hat{\theta}_j}{\hat{\sigma} \sqrt{v_j}}$$

où $v_j = j$ -ème élément sur la diagonale de la matrice $(\mathbf{X}^\top \mathbf{X})^{-1}$.

Proposition 367. Sous l'hypothèse nulle que $\theta_j = 0$, on a

$$z_j \sim t(n - p - 1).$$

où $t(n - p - 1)$ est la loi de Student à $n - p - 1$ degré de liberté (voir définition 270 page 91)

Aller plus loin

Régularisation (cf. poly de Chloé-Agathe Azencott)

Chapitre 17

Régression logistique

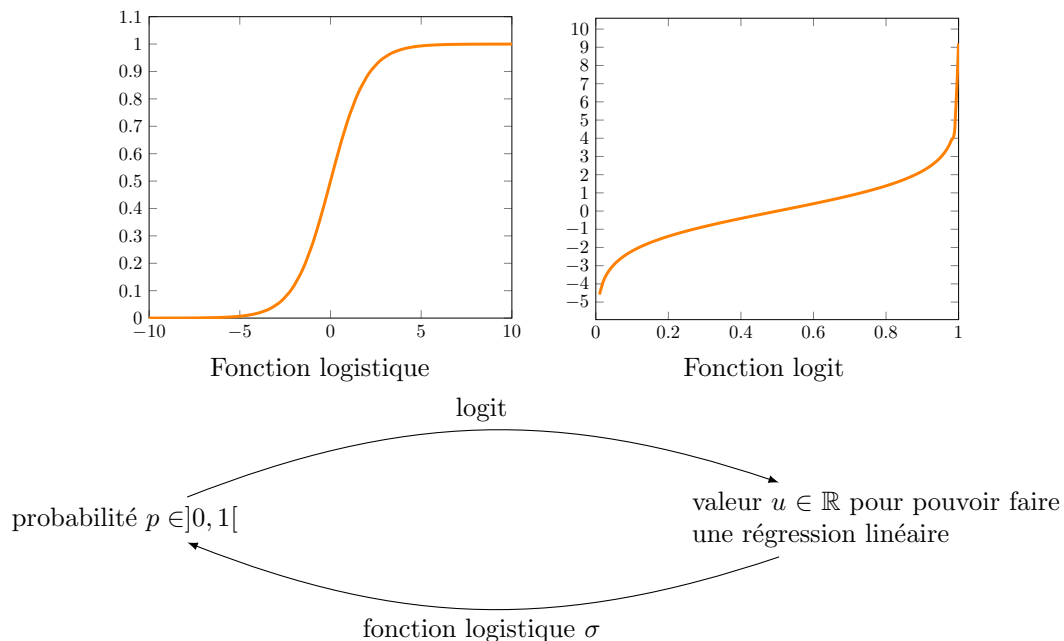
17.1 Idée informelle

Définition 368 (régression binaire). Une régression binaire est un problème qui a la forme suivante :

Régression binaire

entrée : des features $x_1, \dots, x_n \in \mathbb{R}^p$, des étiquettes $y_1, \dots, y_n \in \{0, 1\}$
sortie : une fonction f qui colle au plus près : $y_i \approx f(x_i)$ pour tout i

L'idée de la **régression logistique** est d'utiliser la régression linéaire, mais en utilisant la fonction *logit* définie après, pour convertir une probabilité entre $]0, 1[$ dans \mathbb{R} comme requis par la régression linéaire.



Définition 369 (fonction logistique ou fonction sigmoïde).

$$\begin{aligned} \sigma : \mathbb{R} &\rightarrow]0, 1[\\ u &\mapsto \frac{1}{1+e^{-u}} \end{aligned}$$

Définition 370 (fonction logit).

$$\begin{aligned} \text{logit} :]0, 1[&\rightarrow \mathbb{R} \\ p &\mapsto \ln \frac{p}{1-p} \end{aligned}$$

Notation σ .

C'est ça les statistiques aussi. Utiliser les même notations pour des objets différents. Ne jamais louper une opportunité de rendre les choses floues. Ici σ désigne la fonction logistique, ce n'est pas l'écart-type d'une variable aléatoire. D'ailleurs les statisticiens appellent cette fonction la fonction logistique, mais en fait, il y a des fonctions logistiques (voir Wikipedia).

On considère un couple (X, Y) de variables aléatoires où X à valeurs dans \mathbb{R}^p et Y à valeur dans $\{0, 1\}$. On suppose :

$$\mathbb{P}(Y = 1 \mid X = x) = \sigma(\langle \theta, x \rangle)$$

où $\theta \in \mathbb{R}^p$ est inconnu.

17.2 Exemples d'application

Expliquer la survenue d'un accident vasculaire cérébral en fonction de $x = \begin{pmatrix} \text{taux de glucose} \\ \text{taux de cholestérol} \\ \text{pollution} \\ \text{pratique sportive} \end{pmatrix}$

Expliquer l'achat d'un shampoing en fonction de $x = \begin{pmatrix} \text{position du logo} \\ \text{taille du logo} \\ \text{taille de la police} \\ \text{position en rayon} \end{pmatrix}$

17.3 Définition

Définition 371. Régression logistique

entrée : des features $x_1, \dots, x_n \in \mathbb{R}^p$, des étiquettes $y_1, \dots, y_n \in \{0, 1\}$

sortie : les valeurs des paramètres θ qui maximise la vraisemblance des données $x_1, \dots, x_n, y_1, \dots, y_n$ par rapport au modèle

$$\mathbb{P}(Y = 1 \mid X = x) = \sigma(\langle \theta, x \rangle)$$



Théorème 372. La solution du problème de régression logistique est un θ qui maximise

$$\sum_{i=1}^n y_i \log(\sigma(\langle \theta, x_i \rangle)) + (1 - y_i) \log(1 - \sigma(\langle \theta, x_i \rangle)).$$

DÉMONSTRATION.

La vraisemblance s'écrit :

$$\begin{aligned} \mathcal{L}(x, y ; \theta) &= \prod_{i=1}^n \mathbb{P}(X = x_i, Y = y_i \mid \theta) \\ &= \prod_{i=1}^n \mathbb{P}(Y = y_i \mid X = x_i, \theta) \times \mathbb{P}(X = x_i) \\ &= \text{cte} \times \prod_{i=1}^n \mathbb{P}(Y = y_i \mid X = x_i, \theta) \\ &= \text{cte} \times \prod_{i=1}^n \mathbb{P}(Y = 1 \mid X = x_i, \theta)^{y_i} \mathbb{P}(Y = 0 \mid X = x_i, \theta)^{1-y_i} \end{aligned}$$

Le résultat du théorème c'est la log-vraisemblance où on a enlevé la constante. ■

Aller plus loin

Test de Hosmer–Lemeshow (faire la page Wikipedia par exemple)

Chapitre 18

Classification bayésienne naïve

Exemple 373. Filtre de spam

entrée : un email x

sortie : l'email x est sans doute un spam, ou non, l'email x n'est pas un spam.

Considérons un jeu de données $(x_1, y_1, \dots, x_n, y_n)$ avec $x_1, \dots, x_n \in \mathbb{R}^p$ et $y_1, \dots, y_n \in \{0, 1\}$. Notre modélisation est de considérer que ce jeu de données proviennent d'un couple aléatoire (X, Y) où $X : \Omega \rightarrow \mathbb{R}^p$ et $Y : \Omega \rightarrow \{0, 1\}$. On parle de **modélisation générative**.

Exemple 374 (spams). On considère n emails. le i -ème email est (x_i, y_i) où x_i est une représentation vectorielle du contenu de l'email (de type bag-of-words) et

$$y_i = \begin{cases} 1 & \text{s'il s'agit d'un spam} \\ 0 & \text{sinon.} \end{cases}$$

Plus précisément, on se fixe p mots comme 'riche', 'célèbre', 'argent', etc. Puis

$$x_{ij} = \begin{cases} 1 & \text{si le } j\text{-ème mot apparaît dans le } i\text{-ème email} \\ 0 & \text{sinon.} \end{cases}$$

Prédiction

entrée : une donnée x

sortie : oui si $\mathbb{P}(Y = 1 \mid X = x) > \mathbb{P}(Y = 0 \mid X = x)$

Exemple 375. Un email x , le déclarer comme spam si la probabilité qu'il soit du spam sachant qu'il s'agit de l'email x est plus grande que la probabilité de ne pas être du spam.

18.1 Bayes est gentil

Pour pouvoir prédire il nous faut calculer $\mathbb{P}(Y = 1 \mid X = x)$ et $\mathbb{P}(Y = 0 \mid X = x)$. Pour cela, on utilise la formule de Bayes :

$$\mathbb{P}(Y = 1 \mid X = x) = \frac{\mathbb{P}(X = x \mid Y = 1)\mathbb{P}(Y = 1)}{\mathbb{P}(X = x)}$$

$$\mathbb{P}(Y = 0 \mid X = x) = \frac{\mathbb{P}(X = x \mid Y = 0)\mathbb{P}(Y = 0)}{\mathbb{P}(X = x)}$$

Pour faire la comparaison $\mathbb{P}(Y = 1 \mid X = x) > \mathbb{P}(Y = 0 \mid X = x)$, pas besoin de connaître $\mathbb{P}(X = x)$! On a :

Le problème de prédiction devient donc :

Prédiction

entrée : une donnée x

sortie : oui si $\mathbb{P}(X = x \mid Y = 1)\mathbb{P}(Y = 1) > \mathbb{P}(X = x \mid Y = 0)\mathbb{P}(Y = 0)$

18.2 Naïveté

On note $X = (X_1, \dots, X_p)$. On suppose que les X_i sont indépendants deux à deux conditionnellement à Y . Autrement dit pour tout $j, k \in \{1, \dots, p\}$, $j \neq k$, on a :

$$\mathbb{P}(X_j = v \mid Y = y, X_k = w) = \mathbb{P}(X_j = v \mid Y = y).$$

C'est vraiment naïf. Par exemple le mot ‘célibataire’ a plus de chance d’apparaître s’il y a aussi le mot ‘rencontre’. Ici, on fait comme si les X_j étant indépendants deux à deux. En pratique, le classificateur peut avoir de bons résultats.

Le problème de prédiction devient donc :

Prédiction

entrée : une donnée x

sortie : oui si $\prod_{j=1}^p \mathbb{P}(X_j = x_j \mid Y = 1) \mathbb{P}(Y = 1) > \prod_{j=1}^p \mathbb{P}(X_j = x_j \mid Y = 0) \mathbb{P}(Y = 0)$

18.3 Estimations

Pour vérifier si $\prod_{j=1}^p \mathbb{P}(X_j = x_j \mid Y = 1) \mathbb{P}(Y = 1) > \prod_{j=1}^p \mathbb{P}(X_j = x_j \mid Y = 0) \mathbb{P}(Y = 0)$ ou non, on estime chacun des termes. Plus précisément, on considère un modèle statistique où les paramètres θ sont :

$\mathbb{P}(Y = 1)$	probabilité qu’un email quelconque soit un spam
$\mathbb{P}(X_j = 1 \mid Y = 1)$	probabilité qu’un email contiennent le j -ème mot sachant que c’est un spam
$\mathbb{P}(X_j = 1 \mid Y = 0)$	probabilité qu’un email contiennent le j -ème mot sachant que ce n’est pas un spam

Théorème 376. Soit $\mathcal{D} = (x_i, y_i)_{i=1..n}$. Les estimations par maximum de vraisemblance par rapport aux données \mathcal{D} sont données par :

$$\begin{array}{lll} \mathbb{P}(Y = 1) & \text{estimé par} & \frac{n_{\text{spam}}}{n} \\ \mathbb{P}(X_j = 1 \mid Y = 1) & \text{estimé par} & \frac{\sum_{i=1}^n x_{ij} y_j}{n_{\text{spam}}} \\ \mathbb{P}(X_j = 1 \mid Y = 0) & \text{estimé par} & \frac{\sum_{i=1}^n x_{ij} (1 - y_j)}{n - n_{\text{spam}}} \end{array}$$

$$\text{où } n_{\text{spam}} = \sum_{i=1}^n y_i.$$

18.4 Limite du modèle

Attention, si par exemple, le j -ème mot ‘drosophile’ n’apparaît dans aucun spam, alors il est impossible de prédire que un email sans ce j -ème soit un spam ! Jamais !

On peut utiliser un lissage de Laplace.

Troisième partie

Partie théorique

Chapitre 19

Fonctions caractéristiques

Dans ce chapitre, les démonstrations sont présentées sous la forme d'exercices. Le but est de démontrer le **théorème centrale limite**, qui est une convergence en loi : $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2)$ (Section 19.6). Pour le démontrer, nous allons introduire la notion de **fonction caractéristique** (Section 19.1) qui est une sorte de **transformée de Fourier** d'une loi, et qui la caractérise complètement (Section 19.3). Dans la démonstration du TCL, nous nous appuyons sur les propriétés suivantes : **indépendance** des variables et information sur les **moments** (espérance et variance). Ces propriétés se 'retrouvent' dans les fonctions caractéristiques (Section 19.2). La convergence en loi se démontre grâce à la convergence ponctuelle des fonctions caractéristiques via le **théorème de convergence de Levy** (Section 19.5). La démonstration du théorème de Levy repose sur les notions de **famille tendue de mesures** (Section 19.4).

19.1 Définition

Définition 377. Soit X une variable aléatoire. La **fonction caractéristique** de X est la fonction

$$\varphi_X(t) = \mathbb{E}(e^{itX}).$$

Proposition 378. Si X est à densité p , alors la fonction caractéristique de X est la **transformée de Fourier** de p :

$$\varphi_X(t) = \int_{\mathbb{R}} p(x) e^{itx} dx.$$

La fonction caractéristique ne dépend que de la loi de X et pas de la variable aléatoire X .

Proposition 379. $\varphi_{\mathcal{N}(0,1)}(t) = e^{-\frac{t^2}{2}}$.

19.2 Propriétés

Théorème 380. Si X et Y sont indépendantes alors $\varphi_{X+Y} = \varphi_X \cdot \varphi_Y$.

DÉMONSTRATION.

1. Si X et Y sont indépendantes, est-ce que $f(X)$ et $g(Y)$ sont indépendantes ?
2. Si Z_1 et Z_2 sont indépendantes, que vaut $\mathbb{E}(Z_1 Z_2)$?
3. Conclure.

■

Proposition 381 (lien entre dérivées et moments). Si X admet un moment d'ordre n , alors φ_X est de classe C^n et

$$\text{pour tout } k \in \{1, \dots, n\}, \varphi_X^{(k)}(0) = i^k \mathbb{E}(X^k).$$

DÉMONSTRATION.

1. Rappeler le théorème de dérivabilité sous le signe intégrale.
2. Montrer la proposition par récurrence sur k .

■

19.3 Caractérisation

Théorème 382. Soit X et Y deux variables aléatoires.

$\varphi_X = \varphi_Y$ implique que X et Y suivent la même loi.

DÉMONSTRATION.

1. Montrer que

$$\int_{-T}^T T \frac{e^{i\lambda t}}{it} dt = \text{sign}(\lambda) \int_{-|\lambda|T}^{|\lambda|T} \frac{\sin t}{t} dt.$$

2. Soit μ une loi de probabilité. On pose :

$$I_T(a, b) := \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ta} - e^{-ib}}{it} \varphi_\mu(t) dt.$$

Montrer que :

$$I_T(a, b) = \int_{\mathbb{R}} \left(\frac{\text{sign}(x-a)}{2\pi} \int_{-|x-a|T}^{|x-a|T} \frac{\sin t}{t} dt - \frac{\text{sign}(x-b)}{2\pi} \int_{-|x-b|T}^{|x-b|T} \frac{\sin t}{t} dt \right) d\mu(x).$$

Indice : on admettra le théorème de Fubini pour échanger les signes d'intégration \int .

3. Montrer que

$$I_T(a, b) \xrightarrow{T \rightarrow +\infty} \mu([a, b]) + \frac{1}{2}(\mu(a) + \mu(b)).$$

On utilise le fait que $\int_{-y}^y \frac{\sin t}{t} dt \xrightarrow{y \rightarrow +\infty} \pi$.

4. Supposons $\varphi_X = \varphi_Y$. Notons \mathcal{I} l'ensemble des intervalles $]a, b[$ avec

$$a, b \in \mathbb{R} \setminus \{x \in \mathbb{R} \mid \mathbb{P}_X(\{x\}) > 0 \text{ ou } \mathbb{P}_Y(\{x\}) > 0\}.$$

(a) Conclure que $\mathbb{P}_X([a, b]) = \mathbb{P}_Y([a, b])$ pour tout $]a, b[$ dans \mathcal{I} .

(b) En déduire que $\mathbb{P}_X = \mathbb{P}_Y$.

■

19.4 Famille tendue de mesures

Définition 383. Une famille \mathcal{F} de mesures est **tendue** si pour tout $\epsilon > 0$, il existe un compact K tel que pour tout mesure $\mu \in \mathcal{F}$, $\mu(K) \geq 1 - \epsilon$.

Proposition 384. Une famille finie est tendue.

Définition 385. Une famille \mathcal{F} de mesures est **relativement compacte** si toute suite d'éléments de \mathcal{F} admet une sous-suite qui converge en loi.

Théorème 386 (de Prohorov). Toute famille tendue de mesures est relativement compacte.

DÉMONSTRATION.

Soit \mathcal{F} une famille tendue de mesures. Soit $(\mu_n)_{n \in \mathbb{N}}$ une suite de mesures dans \mathcal{F} . On note F_n la fonction de répartition de μ .

1. On admet le théorème de Helly, il existe une suite extraite $(F_{\lambda_n})_{n \in \mathbb{N}}$ et une fonction F qui est croissante et continue à droite telle que

$$F_{\lambda_n}(x) \rightarrow_{n \rightarrow +\infty} F(x)$$

en tout point x de continuité de F .

2. Montrer que $F(x) \xrightarrow{x \rightarrow -\infty} -1$ et $F(x) \xrightarrow{x \rightarrow +\infty} 1$.

Indice : c'est là que l'on utilise l'hypothèse que \mathcal{F} est **tendue**.

3. En déduire avec le corollaire 131 page 43 que F est la fonction de répartition d'une loi μ .

4. Conclure que $\mu_{\lambda_n} \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mu$.

■

Proposition 387. Soit $(\mu_n)_{n \in \mathbb{N}}$ une suite de mesures de probabilité telle que :

1. $\{\mu_n, n \in \mathbb{N}\}$ est tendue
2. toute sous-suite de $(\mu_n)_{n \in \mathbb{N}}$ qui converge en loi, converge vers μ

Alors $\mu_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mu$.

DÉMONSTRATION.

Nous allons que pour toute fonction f bornée et continue on a

$$u_n := \int_{\mathbb{R}} f(x) d\mu_n(x) \rightarrow_{n \rightarrow +\infty} \int_{\mathbb{R}} f(x) d\mu(x).$$

1. Considérons n'importe quelle sous-suite $(\mu_{\lambda_n})_{n \in \mathbb{N}}$. D'après le théorème de Prokhorov, il existe une sous-suite extraite de cette sous-suite qui converge en loi : $(\mu_{\lambda_{m_n}})_{n \in \mathbb{N}}$. Quelle est la limite de $(\mu_{\lambda_{m_n}})_{n \in \mathbb{N}}$?
2. Soit fonction f bornée et continue. Quelle est la limite de $u_{\lambda_{m_n}}$?
3. En déduire que

$$u_n \rightarrow_{n \rightarrow +\infty} \int_{\mathbb{R}} f(x) d\mu(x).$$

4. Conclure que $\mu_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mu$.

■

19.5 Théorème de Levy

Théorème 388 (de Levy). $X_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} X$ ssi pour tout $t \in \mathbb{R}$, $\varphi_{X_n}(t) \xrightarrow[n \rightarrow \infty]{} \varphi_X(t)$.

DÉMONSTRATION.

— Montrer \Rightarrow

— Montrons \Leftarrow . Supposons $\varphi_{X_n} \rightarrow_{n \rightarrow +\infty} \varphi_X$. On note μ_n la loi de probabilité de X_n sur \mathbb{R} . On note μ la loi de probabilité de X sur \mathbb{R} .

1. Montrons que la famille $\{\mu_n, n \in \mathbb{N}\}$ est **tendue**. Par définition d'être tendue, il faut montrer qu'il existe un compact K' tel que pour tout $n \in \mathbb{N}$, $\mu_n(K') \geq 1 - \epsilon$.
 - (a) Soit $n \geq 0, a \geq 0$. Montrer que $\frac{1}{2a} \int_{-a}^a (1 - \varphi_{\mu_n}(t) dt) \geq (1 - \frac{2}{\pi}) \mu_n(\{x \in \mathbb{R} \mid |x| \geq \frac{\pi}{2a}\})$.
Indice : calcul avec $\frac{\sin ax}{ax}$
 - (b) Montrer qu'il existe n_0 tel que pour tout $n \geq n_0$, $\mu_n(\{x \in \mathbb{R} \mid |x| \geq \frac{\pi}{2a}\}) \leq \epsilon$.
 - (c) Expliquer pourquoi $\{\mu_n, n \leq n_0\}$ est tendue.
 - (d) Il existe K compact avec pour tout $n \leq n_0$, $\mu_n(K) \leq 1 - \epsilon$ On pose $K' := K \cup [-\frac{\pi}{2a}, \frac{\pi}{2a}]$. Montrer que pour tout $n \in \mathbb{N}$, $\mu_n(K') \geq 1 - \epsilon$.
2. Montrer que toute sous-suite de $(\mu_n)_{n \in \mathbb{N}}$ qui converge en loi, converge vers μ .
3. En déduire que $\mu_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mu$. Autrement dit, $X_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} X$.

■

19.6 Démonstration du théorème centrale limite

Théorème 389 (théorème central limite). *Soit $(X_n)_n$ iid admettant un moment d'ordre 2. On note μ l'espérance et σ^2 la variance de X_n . Alors :*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2).$$

Au lieu de démontrer le théorème dans toute sa généralité, nous allons le démontrer dans le cas où l'espérance est nulle et la variance vaut 1.

Théorème 390 (théorème central limite simplifié). *Soit $(X_n)_n$ iid de l'espérance nulle, et de variante 1. Alors :*

$$\sqrt{n} \cdot \bar{X}_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

DÉMONSTRATION.

1. But : montrer que $\varphi_{\sqrt{n}\bar{X}_n}$ converge vers $\varphi_{\mathcal{N}(0,1)}(t)$ point par point.

(a) Montrer que $\varphi_{\sqrt{n}\bar{X}_n}(t) = \varphi_X(\frac{t}{\sqrt{n}})^n$.

Indice : utiliser le fait que les X_1, \dots, X_n soient **indépendantes**

(b) Montrer que $\varphi_X(\frac{t}{\sqrt{n}}) = 1 - \frac{t^2}{2n} + o(1)$.

Indice : utiliser le **lien entre dérivées et moments**

(c) Montrer que $e^{\frac{-t^2}{2n}} = 1 - \frac{t^2}{2n} + o(\frac{1}{n})$.

(d) Soit $z, u \in \mathbb{C}$ avec $|z| \leq 1, |u| \leq 1$. Montrer que $|z^n - u^n| \leq n|z - u|$.

(e) Montrer que $|\varphi_X(\frac{t}{\sqrt{n}})^n - e^{\frac{-t^2}{2}}| = o(1)$.

2. Conclure avec le **théorème de Levy**.

■

Bibliographie

- [Aze22] Chloé-Agathe Azencott. *Introduction au Machine Learning-2e éd.* Dunod, 2022.
- [BL21] Philippe Barbe and Michel Ledoux. Probabilité. In *Probabilité*. EDP Sciences, 2021.
- [GK19] Olivier Garet and Aline Kurtzmann. *De l'intégration aux probabilités-2e édition augmentée*. Editions Ellipses, 2019.
- [HTFF09] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning : data mining, inference, and prediction*, volume 2. Springer, 2009.
- [Lec02] Jean-Pierre Lecoutre. *Statistique et probabilités*. Dunod, 2002.