

Cours de probabilités de L2 mathématique

Yoann Dabrowski

20 mars 2019

Table des matières

1	Ensembles, Cardinaux, Dénombrement	4
1	Ensembles	4
2	Ensembles finis	5
3	Opérations sur les ensembles finis et leurs cardinaux	6
3.1	Compléments	8
3.2	Rappels sur les sommes indicées par des ensembles finis	8
4	Dénombrement	10
4.1	Arrangement	10
4.2	Permutation	11
4.3	Combinaison	12
4.4	Rappels des Propriétés des coefficients $\binom{n}{p}$	12
4.5	Arrangement avec répétition	12
4.6	Combinaison avec répétition (facultatif)	14
5	Application : espaces de probabilités finis	14
5.1	La probabilité uniforme $\mathcal{U}(n)$ et dénombrement en pratique en probabilité	16
5.2	Les variables aléatoires et leurs lois	16
5.3	Moyenne ou Espérance ou Moment d'ordre 1	18
5.4	Quelques lois usuelles sur un ensemble fini	19
2	Statistiques descriptives : Première partie	20
1	Types de variables statistiques et leurs Représentations graphiques	20
1.1	Variable qualitative nominale et représentations de leurs tables de contingence	22
1.2	Variables qualitatives ordinales et diagramme cumulatif	24
1.3	Variables quantitatives discrètes et diagramme bâton	24
1.4	Variables quantitatives continues et histogrammes	26
2	Résumés numériques communs d'une variable statistique quantitative	27
2.1	Moyenne et variance empirique	27
2.2	Médiane et quartiles	28
2.3	Diagrammes à moustaches (ou Diagrammes à boîtes, boxplot)	30
3	Familles sommables	31
1	Complément au chapitre 1 : Ensembles infinis	31
1.1	Ensembles infinis dénombrables	31
1.2	Ensembles infinis non dénombrables	33
2	Familles sommables à termes positifs	34
3	Familles sommables à termes scalaires	36
4	Familles de carrés sommables et espaces $L^p(\Omega, P)$	38

4.1	Familles de carré sommable (facultatif)	39
4.2	Espaces $L^p(\Omega, P)$	39
4	Espaces probabilisés	41
1	Expérience aléatoire, Introduction	41
2	L'espace des réalisations	41
3	Évènements et tribu d'évènements	42
4	Probabilité sur une tribu d'évènements	43
5	Premières propriétés d'une probabilité	43
6	Probabilité sur un ensemble fini ou dénombrable	45
6.1	Cas des ensembles finis	45
6.2	Cas d'un ensemble dénombrable	45
6.3	Quelques lois usuelles sur un ensemble fini ou dénombrable	46
5	Probabilité conditionnelle. Indépendance	47
1	Probabilité conditionnelle	47
2	Indépendance	49
2.1	Retour sur l'interprétation des lois géométriques et binomiales	51
6	Variables aléatoires discrètes	53
1	Introduction	53
2	Exemples	53
3	Définitions dans le cas général. Loi d'une variable aléatoire (facultatif)	54
4	Variables aléatoires discrètes	55
4.1	Loi d'une v.a.d.	56
4.2	Suites de v.a. indépendantes	56
4.3	Fonction de répartition (f.r.) d'une v.a.d.	57
4.4	Moyenne ou Espérance ou Moment d'ordre 1	57
4.5	Théorème de transfert	61
4.6	Moments d'une v.a.d.	62
5	Couples aléatoires (avec de nombreux résultats admis)	63
5.1	Couples de v.a. discrètes	64
6	Compléments sur les variables aléatoires discrètes	65
6.1	Variance et Écart-type d'une v.a.d. Variable centrée réduite	65
6.2	Inégalités de Markov et de Bienaymé-Tchébychev	65
6.3	Fonctions génératrice d'une v.a.d.	66
7	Séries génératrices	68
1	Dérivation de la série génératrice	68
2	Première application en combinatoire : Nombre de solutions de l'équation $\sum_{k=1}^n x_k = p$	69
3	Une Autre Application de la méthode des séries génératrices : nombre de parenthésages	70
3.1	Application au calcul de loi de sommes de v.a. indépendantes	71
8	Théorème de Moivre-Laplace I : Cadre et Enoncé	72
1	Introduction aux Variables aléatoires continues (ou à densité)	72
1.1	Densité d'une probabilité	72
1.2	Lois à densité classiques	73
1.3	Théorème de transfert (admis)	74

1.4	Fonction de répartition (f.r.) d'une v.a.c. (facultatif)	74
1.5	Moyenne ou Espérance ou Moment d'ordre 1 (facultatif)	74
1.6	Variance et Écart-type d'une v.a.c. Variable centrée réduite (facultatif)	75
2	Théorème de Moivre-Laplace proprement dit	76
3	Autres théorèmes limites (facultatif et seulement les 2 premiers énoncés vus en cours)	76
3.1	Fonction caractéristiques	77
3.2	Idées de preuves	78
9	Statistiques descriptives : Deuxième Partie	81
1	Représentations graphiques bivariées et analyse en composantes principales	81
2	Introduction à la statistique inférentielle	84
2.1	Exemples d'intervalles de confiance pour le paramètre d'une variable de Bernoulli	84
2.2	Un exemple de test d'hypothèse : le test d'indépendance du χ^2	85
10	Théorème de Moivre-Laplace II : Démonstration	89
1	Formule de Stirling	89
2	Théorème de Moivre-Laplace proprement dit	90

Chapitre 1

Ensembles, Cardinaux, Dénombrément

1 Ensembles

Un ensemble Ω est une collection d'objets appelés éléments de cet ensemble. On écrit $\omega \in \Omega$ pour exprimer que ω est un élément de Ω : C'est la relation d'appartenance. Deux éléments ω et ω' de Ω , sont égaux et on écrit $\omega = \omega'$, ou différents et on écrit $\omega \neq \omega'$. L'ensemble vide, noté \emptyset ne contient aucun élément.

Définition 1. Un ensemble A est appelé sous-ensemble ou partie de Ω si tout élément de A est élément de Ω . On écrit $A \subset \Omega$ pour exprimer que A est une partie de Ω : C'est la relation d'inclusion. Un élément $\omega \in \Omega$ peut appartenir ou ne pas appartenir à la partie A . On écrit alors respectivement $\omega \in A$ ou $\omega \notin A$. Le complémentaire de A , noté A^c , est l'ensemble des éléments $x \in \Omega$ tels que $x \notin A$.

Définition 2. La fonction indicatrice d'une partie A est l'application $1_A : \Omega \rightarrow \{0; 1\}$ définie par $1_A(\omega) = 1$ si $\omega \in A$ et $1_A(\omega) = 0$ si $\omega \notin A$.

On a admis en L1 l'existence de l'ensemble \mathbb{N} des entiers naturels et d'un ensemble constitué des parties de Ω (ce sont des axiomes de base de la théorie des ensembles).

Définition 3. L'ensemble des parties de Ω est noté $\mathcal{P}(\Omega)$. Une famille F de parties de Ω est une partie de $\mathcal{P}(\Omega)$. Les éléments de F sont des parties de Ω .

Lemme 1. La fonction indicatrice $A \mapsto 1_A$ réalise une bijection entre $\mathcal{P}(\Omega)$ et $\{0; 1\}^\Omega$.

Démonstration. L'inverse est $h \mapsto h^{-1}(\{1\})$. La vérification que c'est bien un inverse est facile, et laissée en exercice. \square

Remarque 1. Si A et B sont deux parties de Ω (i.e. deux éléments de $\mathcal{P}(\Omega)$).

1. On a les relations $A \subset B$ ou $B \subset A$ ou ($A \not\subset B$ et $B \not\subset A$). $A \subset B$ s'écrit aussi $B \supset A$.
2. On a défini en L1 : $A \times B$ l'ensemble des couples (a, b) $a \in A, b \in B$, l'intersection $A \cap B$ (ensemble des éléments à la fois dans A **et** dans B), l'union $A \cup B$ (ensemble des éléments à la fois dans A **ou** dans B), le complémentaire de B dans A : $A - B = A \cap B^c = \{x \in A : x \notin B\}$ et la différence symétrique $A \Delta B = (A - B) \cup (B - A)$. On remarquera la relation de ces opérations avec les connecteurs logiques de base.
3. Plus généralement on définit l'union d'une famille $A_i \in \mathcal{P}(\Omega), i \in I$:

$$\bigcup_{i \in I} A_i = \{x \in \Omega : \exists i \in I : x \in A_i\},$$

et de l'intersection d'une même famille finie :

$$\bigcap_{i \in I} A_i = \{x \in \Omega : \forall i \in I : x \in A_i\}.$$

4. A et B sont disjoints si $A \cap B = \emptyset$.

2 Ensembles finis

On va définir les ensembles finis comme les ensembles en bijection avec une partie de la forme

$$\llbracket 1, n \rrbracket := \{k \in \mathbb{N} : 1 \leq k \leq n\}$$

pour $n \in \mathbb{N}$. On remarquera que pour $n = 0$, on a $\llbracket 1, n \rrbracket = \emptyset$.

On rappelle que \mathfrak{S}_n est l'ensemble des bijections de $\llbracket 1, n \rrbracket$ dans lui-même.

On commence par un lemme qui dit que ces ensembles ne sont pas en bijection.

Lemme 2. Soient $m, n \in \mathbb{N}$. Si $\llbracket 1, n \rrbracket$ est en bijection avec $\llbracket 1, m \rrbracket$ alors $m = n$.

Démonstration. On montre la propriété par récurrence sur n . Pour $n = 0$, l'énoncé est vrai car l'ensemble vide n'est en bijection qu'avec l'ensemble vide.

Supposons le résultat vrai au rang $n \geq 0$. Soit $h : \llbracket 1, n+1 \rrbracket \rightarrow \llbracket 1, m \rrbracket$ une bijection (donc forcément $m \geq 1$ comme le premier ensemble n'est pas vide). On distingue 2 cas :

Le cas simple est le cas $h(n+1) = m$, alors on définit $g : \llbracket 1, n \rrbracket \rightarrow \llbracket 1, m-1 \rrbracket$ par $g(k) = h(k)$. g est bien défini car h est une bijection donc pour $k \neq n+1$, $h(k) \neq h(n+1) = m$, donc $h(k) < m$. g est injective car la restriction d'une application injective est encore injective. Soit $l \in \llbracket 1, m-1 \rrbracket$ il existe $k \in \llbracket 1, n+1 \rrbracket$ tel que $h(k) = l$ par surjectivité de h . Mais $l \neq n+1$ par injectivité comme ci-dessus, donc $k \in \llbracket 1, n \rrbracket$ et donc $g(k) = l$. Cela montre la surjectivité de g . Donc par l'hyp de récurrence $n = m-1$ et donc $m = n+1$.

Le second cas est le cas $h(n+1) < m$. On a vu au premier semestre l'existence de la transposition $\tau = (h(n+1), m)$ de sorte que $\tau \circ h$ est encore une bijection par composée de bijections et elle vérifie le premier cas. D'où l'étape suivante de la récurrence dans tous les cas. □

Définition 4. On dit qu'un ensemble A est *fini* s'il existe un entier $n \in \mathbb{N}$, qui est alors unique (par le lemme précédent), et une bijection $f : \llbracket 1, n \rrbracket \rightarrow A$. On écrit alors $\text{Card}(A) = n$ ou $|A| = n$. Un ensemble qui n'est pas fini est dit *infini*.

Ainsi, pour un ensemble fini A , l'entier $n = \text{Card}(A)$ représente le nombre d'éléments de A et la bijection $f : \llbracket 1, n \rrbracket \rightarrow A$ permet de représenter les éléments de A à l'aide d'une suite finie et d'écrire $A = \{x_k; 1 \leq k \leq n\}$. ($x_k = f(k)$).

Exemple 1. Soient $p, n \in \mathbb{N}$. L'ensemble

$$\llbracket p+1, p+n \rrbracket := \{k \in \mathbb{N} : p+1 \leq k \leq p+n\}$$

est un ensemble fini de cardinal n (via la bijection $k \mapsto k+p$ de $\llbracket 1, n \rrbracket \rightarrow \llbracket p+1, p+n \rrbracket$.)

Pour raisonner par récurrence sur le cardinal d'un ensemble, le résultat suivant est pratique :

Lemme 3. Si E est un ensemble fini de cardinal $n > 0$ (donc E non vide) et si $a \in E$ alors $E - \{a\}$ est de cardinal $n - 1$.

Démonstration. Si $h : \llbracket 1, n \rrbracket \rightarrow E$ est une bijection. On a 2 cas. Soit $h(n) = a$ et alors la restriction de h à $\llbracket 1, n-1 \rrbracket$ est une bijection sur $E - \{a\}$.

Soit il existe $p < n$ avec $h(p) = a$ (par surjectivité). On considère $\tau = (p, n) \in \mathfrak{S}_n$ la transposition, $g = h \circ \tau$ est une bijection avec $g(n) = h(p) = a$ et cela nous ramène au premier cas. \square

3 Opérations sur les ensembles finis et leurs cardinaux

Proposition 4. Soit E un ensemble fini et $F \in \mathcal{P}(E)$ une partie de E .

1. Alors F est finie et $\text{Card}(F) \leq \text{Card}(E)$.
2. $\text{Card}(F) = \text{Card}(E)$ si et seulement si $E = F$.

En particulier, toute intersection d'ensembles finis est un ensemble fini (comme partie de n'importe lequel des ensembles de l'intersection).

Remarque 2. Un ensemble fini n'est donc pas en bijection avec une de ses parties propres. Par contre \mathbb{N} est en bijection avec les entiers pairs, le résultat n'est donc plus vrai pour un ensemble infini. On ne peut plus dire qu'une partie stricte d'un ensemble infini est plus petite que le tout.

Démonstration. On prouve le résultat par récurrence sur le cardinal de E .

Si $\text{Card}(E) = 0$, E est vide donc F est vide et l'équivalence est aussi évidente.

Supposons le résultat vrai au rang n et prenons E de cardinal $n+1$.

Si $F = E$, F est fini et de cardinal $n+1 \leq n+1$ Sinon il existe $a \in E - F$ donc $F \in E - \{a\}$ et par le lemme 3, cet ensemble a cardinal n et l'hypothèse de récurrence conclut à $\text{Card}(F) \leq n$. Dans les 2 cas, les 2 énoncés de la proposition sont vrais. \square

Proposition 5. Une réunion finie d'ensembles finis est finie et l'on a $\text{Card}(\cup_{k=1}^n A_k) \leq \sum_{k=1}^n \text{Card}(A_k)$, l'égalité ayant lieu si et seulement si les ensembles A_k sont deux à deux disjoints. Si A et B sont finis,

$$\text{Card}(A \cup B) = \text{Card}(A) + \text{Card}(B) - \text{Card}(A \cap B).$$

Démonstration. 1. On traite d'abord le cas d'ensembles disjoints où on veut montrer que $\text{Card}(\cup_{k=1}^n A_k) = \sum_{k=1}^n \text{Card}(A_k)$. Une récurrence simple ramène au cas $n = 2$. Notons $a_i = \text{Card}(A_i)$ Par définition, on a une bijection $g_1 : \llbracket 1, a_1 \rrbracket \rightarrow A_1$. En composant avec la bijection de l'exemple 1, on a une bijection $g_2 : \llbracket 1 + a_1, a_1 + a_2 \rrbracket \rightarrow \llbracket 1, a_2 \rrbracket \rightarrow A_2$. On définit $g : \llbracket 1, a_1 + a_2 \rrbracket \rightarrow A_1 \cup A_2$ par $g(x) = g_1(x)$ pour $x \leq a_1$ et $g(x) = g_2(x)$ pour $x > a_1$. $\text{Im}(g) = \text{Im}(g_1) \cup \text{Im}(g_2) = A_1 \cup A_2$ donc g surjective.

De plus, si $g(x) = g(y)$ alors si la valeur est dans A_i , par définition, on a forcément x, y dans le même cas de définition et $g_i(x) = g_i(y)$, donc $x = y$ par injectivité de g_i . Donc, g est injective et donc bijective. Ceci conclut $\text{Card}(A_1 \cup A_2) = \text{Card}(A_1) + \text{Card}(A_2)$ par définition des cardinaux.

2. Pour le cas A, B finis non-disjoints. On considère les décompositions en ensembles disjoints : $A = (A - B) \cup (A \cap B)$, $B = (B - A) \cup (A \cap B)$ et $A \cup B = (B - A) \cup (A \cap B) \cup (A - B)$ (correspondant à la relation logique $x \in A$ ou $x \in B$ si et seulement si ($x \in A$ et $x \notin B$) ou ($x \in A$ et $x \in B$) ou ($x \in B$ et $x \notin A$)).

On a donc par le cas des ensembles disjoints :

$$\text{Card}(A) = \text{Card}(A - B) + \text{Card}(A \cap B), \quad (a)$$

$$\text{Card}(B) = \text{Card}(B - A) + \text{Card}(A \cap B), \quad (b)$$

$$\text{Card}(A \cup B) = \text{Card}(B - A) + \text{Card}(A \cap B) + \text{Card}(A - B), \quad (c)$$

et donc, en sommant (c)-(a)-(b), on obtient :

$$\text{Card}(A \cup B) = \text{Card}(B - A) + 2\text{Card}(A \cap B) + \text{Card}(A - B) - \text{Card}(A \cap B) = \text{Card}(A) + \text{Card}(B) - \text{Card}(A \cap B)$$

3. On montre l'inégalité dans le cas général. En considérant les ensembles $B_i = A_i \times \{i\} \subset \Omega \times \llbracket 1, n \rrbracket$, on réalise une injection $g : A = \cup_{k=1}^n A_k \rightarrow B = \cup_{k=1}^n B_k$. en posant $g(x) = (x, i)$ avec $i = \min\{j, x \in A_j\}$.

C'est bien une application car i existe toujours, g injective car $p_1 \circ g = id_A$. Donc g réalise une bijection de A sur $Im(g) \subset B$ donc

$$\text{Card}(A) = \text{Card}(Im(g)) \leq \text{Card}(B) = \sum_{k=1}^n \text{Card}(B_k) = \sum_{k=1}^n \text{Card}(A_k)$$

par le cas disjoint et la bijection évidente de A_i, B_i .

4. On montre enfin l'équivalence dans le cas général. Par la proposition 4, l'égalité implique que $Im(g) = B$ dans le point précédent. Si il y avait $x \in A_i \cap A_j$ avec $i < j$ alors (x, j) ne serait pas atteint dans B_j contrairement à la surjectivité. Donc aucun tel x n'existe et les A_i sont deux à deux disjoints. □

Proposition 6. (*Principe du berger*). Soit $f : A \rightarrow B$ une application d'un ensemble fini A vers un ensemble fini B . On suppose qu'il existe $m \in \mathbb{N}^*$ tel que, pour tout $b \in B$; $\text{Card}(f^{-1}(b)) = m$. Alors, $\text{Card}(A) = m\text{Card}(B)$.

Démonstration. Les $f^{-1}(b)$ forment une partition finie de A et en indiquant B par une bijection $B = \{b_1, \dots, b_n\}$, $A = \cup_{k=1}^n f^{-1}(b_k)$ union disjointe (car f application). Donc on a par le calcul du cardinal des unions disjointes : $\text{Card}(A) = \sum_{k=1}^n \text{Card}(f^{-1}(b_k)) = \sum_{k=1}^n m = nm$ comme voulu. □

Le résultat suivant que vous verrez en TD traite le cas d'ensembles non disjoints.

Proposition 7. Plus généralement, si $(A_k)_{1 \leq k \leq n}$ est une suite finie d'ensembles finis, on a la relation suivante

$$\text{Card}(\cup_{k=1}^n A_k) = \sum_{k=1}^n (-1)^{k-1} \sum_{1 \leq i_1 < \dots < i_k \leq n} \text{Card}(A_{i_1} \cap \dots \cap A_{i_k}).$$

Cette relation est connue sous le nom de Formule du crible ou de Poincaré ou encore Principe d'inclusion-exclusion.

Exemple 2. Faire l'exemple à 3 termes ($k=3$). Sur 30 clients ayant commandé un article $A \cup B$, 20 l'ont fait par écrit (A) et 17 l'ont fait par téléphone (B). Le nombre de ceux qui ont utilisé les deux moyens est $|A \cap B| = |A| + |B| - |A \cup B| = 20 + 17 - 30 = 7$.

Proposition 8. Un produit fini d'ensembles finis est fini. Si A et B sont finis, $\text{Card}(A \times B) = \text{Card}(A)\text{Card}(B)$. Plus généralement, si $(A_k)_{1 \leq k \leq n}$ est une suite finie d'ensembles finis, on a

$$\text{Card}(\prod_{k=1}^n A_k) = \prod_{k=1}^n \text{Card}(A_k).$$

Démonstration. Par une récurrence simple, il suffit de montrer le premier calcul (cas $n = 2$). On regarde $p_1 : A \times B \rightarrow A$ la première projection auquel on applique le principe du berger vu $p_1^{-1}(a) = \{a\} \times B$ est en bijection avec B par $b \mapsto (a, b)$. □

L'ensemble des applications d'un ensemble A vers un ensemble B est souvent noté B^A .

Proposition 9. Si A et B sont finis, il en est de même de B^A et l'on a

$$\text{Card}(B^A) = \text{Card}(B)^{\text{Card}(A)}.$$

Démonstration. Si $n = \text{Card}(A)$, on fixe $A = \{a_1, \dots, a_n\}$ une énumération, alors $g : B^A \rightarrow B^n$ donné par $g(f) = (f(a_1), \dots, f(a_n))$ est une bijection qui conclut par le cas des produits. En effet, g bijective car une application est fixée uniquement par ses valeurs. \square

Exemple 3. Le nombre de délégations de 3 membres comprenant un représentant de chacun des trois groupes de 10 ; 15 et 13 étudiants est $10 \cdot 15 \cdot 13 = 1950$.

Exemple 4. On lance un dé à 6 faces puis une pièce de monnaie. Le nombre de résultats possibles est $6 \cdot 2 = 12$.

Exemple 5. En considérant la bijection entre les parties d'un ensemble et leurs fonctions indicatrices, on voit que le nombre de parties d'un ensemble à n éléments est 2^n .

3.1 Compléments

En plus de l'inégalité $\text{Card}(F) \leq \text{Card}(E)$ si il existe une injection $f : F \rightarrow E$, le résultat suivant présente le cas complémentaire des surjections :

Proposition 10. Si $g : E \rightarrow F$ est une surjection et E est fini, alors F est fini et

$$\text{Card}(F) \leq \text{Card}(E).$$

Démonstration. Soit $\sigma : \llbracket 1, n \rrbracket \rightarrow E$ une bijection, on considère $h = g \circ \sigma : \llbracket 1, n \rrbracket \rightarrow F$ surjection et on construit $f : F \rightarrow \llbracket 1, n \rrbracket$ injection par la formule

$$f(x) = \min(h^{-1}(\{x\})).$$

Comme h surjective $h^{-1}(\{x\}) \subset \llbracket 1, n \rrbracket$ donc son minimum est dans $\llbracket 1, n \rrbracket$ de sorte que $f : F \rightarrow \llbracket 1, n \rrbracket$ est bien définie. Si $x \neq y$, on a $h^{-1}(\{x\}) \cap h^{-1}(\{y\}) = \emptyset$ comme pour toute application donc le minimum (qui appartient à chacun de ces ensembles $f(x) \in h^{-1}(\{x\})$) vérifie $f(x) \neq f(y)$. Donc f est injective. Le lemme 4 conclut car

$$\text{Card}(F) = \text{Card}(f(F)) \leq \text{Card}(\llbracket 1, n \rrbracket) = n = \text{Card}(E).$$

\square

3.2 Rappels sur les sommes indicées par des ensembles finis

Si $(a_i)_{i=1, \dots, n}$ est une suite finie (disons de nombres complexes) et $\sigma : \llbracket 1, n \rrbracket \rightarrow \llbracket 1, n \rrbracket$ une bijection. La propriété de commutativité de la somme donne :

$$\sum_{i=1}^n a_i = \sum_{i=1}^n a_{\sigma(i)}.$$

Démonstration. En voyant σ comme produit de transpositions, il suffit de montrer le résultat pour $\sigma = (jk)$ une transposition avec $j < k$.

Mais par commutativité ($a + b = b + a$) et associativité ($(a + b) + c = a + (b + c)$) de la somme :

$$\sum_{i=1}^n a_{\sigma(i)} = \sum_{i=1}^{j-1} a_{\sigma(i)} + a_{\sigma(j)} + \sum_{i=j+1}^{k-1} a_{\sigma(i)} + a_{\sigma(k)} + \sum_{i=k+1}^n a_{\sigma(i)} = \sum_{i=1}^{j-1} a_i + a_k + \sum_{i=j+1}^{k-1} a_i + a_j + \sum_{i=k+1}^n a_i = \sum_{i=1}^n a_i.$$

□

Corollaire 11. Si E est fini et $e : \llbracket 1, n \rrbracket \rightarrow E$ une bijection, $f : E \rightarrow \mathbb{C}$ alors $\sum_{i=1}^n f(e_i)$ ne dépend pas de la bijection e . On note

$$\sum_{e \in E} f(e) = \sum_{i=1}^n f(e_i).$$

Démonstration. Si on prend une autre bijection e' on considère la bijection $\sigma = e^{-1} \circ e'$ de sorte que $e \circ \sigma = e'$. La formule de commutativité de la somme conclut :

$$\sum_{i=1}^n f(e_i) = \sum_{i=1}^n f(e_{\sigma(i)}) = \sum_{i=1}^n f(e'_i).$$

□

Le résultat suivant résume les propriétés de manipulation de ces sommes :

Proposition 12. 1. Si E fini, on a

$$\text{Card}(E) = \sum_{e \in E} 1.$$

2. (Somme par paquet) Si E fini est une union disjointe $E = \bigcup_{i \in I} E_i$ et $f : E \rightarrow \mathbb{C}$ alors :

$$\sum_{e \in E} f(e) = \sum_{i \in I} \sum_{e \in E_i} f(e).$$

3. (interversion de sommes finies) Si E, F sont finis et $a : E \times F \rightarrow \mathbb{C}$, alors :

$$\sum_{e \in E} \sum_{f \in F} a_{e,f} = \sum_{(e,f) \in E \times F} a_{e,f} = \sum_{f \in F} \sum_{e \in E} a_{e,f}.$$

Démonstration. 1. Si $\text{Card}(E) = n$, $E = \{e_1, \dots, e_n\}$, $\sum_{e \in E} 1 = \sum_{i=1}^n 1 = n$ par définition.

2. On suit la preuve pour le cardinal de l'union disjointe. On pose $j : \llbracket 1, m \rrbracket \rightarrow I$ une bijection et $n_i = \text{Card}(E_i)$ On note $N_0 = 0, N_i = \sum_{l=1}^i n_l$.

On a $N_i - N_{i-1} = n_i, i \geq 1$ donc on a une bijection

$$g_i : \llbracket N_{i-1} + 1, N_i \rrbracket \rightarrow E_{j(i)}.$$

Comme au lemme des unions disjointes : on pose $g(k) = g_i(k)$, si $k \in \llbracket N_{i-1} + 1, N_i \rrbracket$ de sorte que g réalise une bijection de $\llbracket 1, N_m \rrbracket \rightarrow E$.

Donc par définition de la somme sur un ensemble (au début et aux deux dernières lignes) :

$$\begin{aligned}
\sum_{e \in E} f(e) &= \sum_{k=1}^{N_m} f(g(k)) \\
&= \sum_{k=1}^{N_1} f(g(k)) + \sum_{k=N_1+1}^{N_2} f(g(k)) + \cdots + \sum_{k=N_{m-1}+1}^{N_m} f(g(k)) \\
&= \sum_{l=1}^m \sum_{k=N_{l-1}+1}^{N_l} f(g(k)) \\
&= \sum_{l=1}^m \sum_{k=N_{l-1}+1}^{N_l} f(g_l(k)) \\
&= \sum_{l=1}^m \sum_{e \in E_j(l)} f(e) \\
&= \sum_{i \in I} \sum_{e \in E_i} f(e)
\end{aligned}$$

3. Il suffit d'appliquer la sommation par paquet aux unions disjointes

$$E \times F = \cup_{e \in E} \{e\} \times F = \cup_{f \in F} E \times \{f\}.$$

□

4 Dénombrement

On appelle dénombrement (ou combinatoire) l'ensemble des techniques qui permettent de déterminer le cardinal d'une partie finie ou d'une famille finie de parties ou d'applications ayant une propriété donnée.

4.1 Arrangement

Définition 5. On appelle *arrangement* de p éléments parmi n , toute suite ordonnée de longueur p (ou p -liste) dont les termes sont des éléments *distincts* d'un ensemble Ω de cardinal n . En d'autres termes, un tel arrangement est une injection $f : \llbracket 1, p \rrbracket \rightarrow \Omega$. Le nombre de ces arrangements, est noté A_n^p .

On note $\text{Inj}(E, F)$ l'ensemble des injections de $E \rightarrow F$. On pose $A_n^p = 0$ si $p > n$ ou si $p < 0$.

Lemme 13. *Le nombre d'arrangements est :*

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

Démonstration. D'abord pour tout n , $A_n^0 = 1$ car il y a une seule injection définie sur l'ensemble vide, l'application vide. On montre la formule par récurrence sur p (simultanément pour tout n). Si $p = 1$, le nombre d'injections de $\{1\}$ dans une ensemble Ω est bien le nombre d'éléments de Ω soit $A_n^1 = n$.

Si on suppose le résultat au rang p , on le montre au rang $p + 1$. A Une injection $f : \llbracket 1, p + 1 \rrbracket \rightarrow \llbracket 1, n \rrbracket$, on associe un couple

$$I(f) = (f(p + 1), (f(p + 1), n) \circ g) \in \Omega \times \text{Inj}(\llbracket 1, p \rrbracket, \llbracket 1, n - 1 \rrbracket)$$

(avec g la restriction de f au p premiers éléments) en effet en composant avec la transposition $(f(p + 1), n)$ qui est bijective on a encore une injection. Donc I est bien défini (l'espace de valeur vient de l'injectivité de f qui force $I(f)_2$ à ne pas prendre 2 fois la même valeur.

I est une bijection d'inverse $(a, h) \mapsto f$ défini par $f(x) = (a, n) \circ h(x)$, $f(p + 1) = a$. Le f ainsi défini est bien injective car si $x \neq y, x, y < p + 1$ $f(x) \neq f(y)$ vient de l'injectivité de h et de la transposition. Et $f(x) \neq f(p + 1)$ vient du choix de la transposition qui ne change pas la valeur de $f(x) = h(x)$ si ce n'est pas a mais remplace $h(x) = a$ par $f(x) = n$ de sorte que a n'est jamais atteint par f sur les p premiers éléments. Vérifier que l'application ci-dessus est bien l'inverse de I est facile.

En conséquence par le cardinal des produits, on a la relation de récurrence : $A_n^p = n \cdot A_{n-1}^{p-1}$ qui donne bien la formule en insérant l'hypothèse de récurrence. \square

Remarque 3. En pratique pour les dénombrements simples on n'explique pas forcément la bijection et on compte juste le nombre de $f(p + 1)$ possible et le nombre d'injections restreintes possibles.

On considère des arrangements quand on parle de "tirage ordonné sans remise" ou "tirage ordonné d'éléments distincts".

Exemple 6. Le nombre de codes à 4 chiffres distincts est $A_{10}^4 = 10 \times 9 \times 8 \times 7 = 5040$.

Exemple 7. Le nombre de bureaux composés d'un président, d'un secrétaire et d'un trésorier qu'une assemblée de 17 personnes peut élire est $A_{17}^3 = 17!/14! = 17 \times 16 \times 15 = 4080$.

4.2 Permutation

Définition 6. On appelle permutation des éléments d'un ensemble Ω de cardinal n toute suite ordonnée de longueur n (ou n -liste) dont les termes sont les éléments distincts de Ω . En d'autres termes, une permutation est une bijection $f : \llbracket 1, n \rrbracket \rightarrow \Omega$. Le nombre de ces permutations est $A_n^n = n!$.

On rappelle qu'on note \mathfrak{S}_n l'ensemble des permutations de $\llbracket 1, n \rrbracket$.

Exemple 8. Sur une carte, le nombre de façons de colorier 6 régions avec 6 couleurs est $6! = 720$.

Exemple 9. Une urne contient 7 jetons numérotés de 1 à 7. Ces jetons sont tirés successivement sans remise. Le nombre de tirages possibles est $7! = 5040$.

En pratique la croissance rapide de $n!$ rend utile d'obtenir un équivalent pour n grand. C'est aussi essentiel pour différents résultats de limites qu'on verra en exercice ou en cours. On rappelle la notion d'équivalent. $u_n \sim v_n$ si $u_n/v_n \rightarrow_{n \rightarrow \infty} 1$. L'équivalent suivant est la base de nombreux résultats asymptotiques en combinatoire ou probabilités, nous verrons la preuve au chapitre 8 :

Lemme 14. (*Formule de Stirling*)

$$n! \sim \sqrt{2\pi n} e^{-n} n^n.$$

4.3 Combinaison

Définition 7. On appelle *combinaison de p éléments parmi n* (ou de n éléments de pris p à p), tout ensemble de cardinal p dont les éléments appartiennent à un ensemble Ω de cardinal n . En d'autres termes, une telle combinaison est une partie à p éléments de Ω (on note $\mathcal{P}_p(\Omega)$ l'ensemble de ces parties). Le nombre de ces combinaisons est noté C_n^p ou $\binom{n}{p}$.

On pose $\binom{n}{p} = 0$ si $p > n$ ou si $p < 0$.

Lemme 15. Le nombre de combinaison est :

$$\binom{n}{p} = \frac{n!}{p!(n-p)!} = \frac{n(n-1)\cdots(n-p+1)}{p!}.$$

Démonstration. On utilise le principe du berger. A une injection de $f \in \text{Inj}(\llbracket 1, p \rrbracket, \llbracket 1, n \rrbracket)$ on associe $\text{Im}(f) \in \mathcal{P}_p(\llbracket 1, n \rrbracket)$. L'image a cardinal p car pour une injection, elle est en bijection avec le domaine de l'application. A chaque partie, on a une unique injection croissante atteignant cette partie et on peut permuter en composant avec $p!$ permutations du domaine, le principe du berger donne donc $A_n^p = p! \binom{n}{p}$ comme voulu. \square

On considère des combinaisons quand on parle de "tirage non-ordonné sans remise" ou "tirage non-ordonné d'éléments distincts".

Exemple 10. Le nombre de listes classées par ordre alphabétique de 4 noms parmi 11 est $\binom{11}{4} = \frac{11 \times 10 \times 9 \times 8}{4!} = 330$.

Exemple 11. Le nombre de choix possibles pour désigner 5 places visiteurs dans un parking de 31 places est $\binom{31}{5} = \frac{31!}{5!(26)!} = 169911$.

4.4 Rappels des Propriétés des coefficients $\binom{n}{p}$

Les propriétés suivantes ont été vues en L1.

Proposition 16. *Formule du binôme de Newton.* $(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}$, x et y des nombres complexes (ou deux éléments d'un anneau unitaire tel que $xy = yx$).

En raison de cette formule, les nombres $\binom{n}{k}$ sont appelés *coefficients du binôme*.

Proposition 17. On a les relations : $\binom{n}{0} = \binom{n}{n} = 1$, $\binom{n}{p} = \binom{n}{n-p}$, $p \binom{n}{p} = n \binom{n-1}{p-1}$, et le triangle de Pascal.

$$\binom{n}{p} = \binom{n-1}{p} + \binom{n-1}{p-1}.$$

4.5 Arrangement avec répétition

Définition 8. Soit $\Omega = \{a_1, \dots, a_n\}$ un ensemble à n éléments, p un entier et $(r_1; \dots; r_n)$ une suite d'entiers telle que $\sum_{k=1}^n r_k = p$. On appelle *arrangement avec répétition*, de longueur p et d'ordre $(r_1; \dots; r_n)$ toute suite ordonnée de longueur p (ou p -liste) dont les termes sont des éléments de Ω où, pour tout $k \in \llbracket 1, n \rrbracket$, chaque a_k figure r_k fois. En d'autres termes, un tel arrangement est une application $f : \llbracket 1, p \rrbracket \rightarrow \Omega$ telle que pour tout $k \in \llbracket 1, n \rrbracket$, $\text{Card}(f^{-1}(a_k)) = r_k$. C'est aussi le nombre de p -tirages avec remise dans une population de taille n , pour lesquels le même individu de la population apparaît r_k fois en tenant compte de l'ordre d'apparition.

Proposition 18. *Le nombre des ces arrangements avec répétition, de longueur p et d'ordre $(r_1; \dots; r_n)$ est $\frac{p!}{r_1! \dots r_n!}$*

Démonstration. L'arrangement avec répétition de référence est a donné par $a(1) = \dots = a(r_1) = a_1, a(r_1 + \dots + r_i + 1) = \dots = a(r_1 + \dots + r_{i+1}) = a_{i+1}$. A une permutation $g \in \mathfrak{S}_p$, on associe un arrangement avec répétition $f = a \circ g$. Pour appliquer le principe des bergers, il suffit de noter que l'ensemble des g tels que $f = a \circ g$ s'identifie à l'ensemble des permutations laissant stable $\llbracket r_1 + \dots + r_i + 1, r_1 + \dots + r_{i+1} \rrbracket$ pour tout i , soit $\mathfrak{S}_{r_1} \times \dots \times \mathfrak{S}_{r_n}$ (en fixant g_0 dans l'ensemble, l'identification est $g \mapsto g_0^{-1}g$). Son cardinal est le dénominateur, d'où le résultat. \square

Vu la ressemblance avec les coefficients binômiaux, ces nombres s'appellent coefficient multinomiaux et sont notés :

$$\binom{p}{r_1, \dots, r_n} = \frac{p!}{r_1! \dots r_n!}.$$

Le nombre total de tous les arrangements avec répétition de p objets parmi n (soit de toutes les p -listes) est n^p (c'est le cardinal de l'ensemble des applications déjà vu plus haut).

Plus généralement, on a la formule du multinôme :

$$\sum_{r_1 + \dots + r_n = p, r_i \geq 0} \frac{p!}{r_1! \dots r_n!} x_1^{r_1} \dots x_n^{r_n} = (x_1 + \dots + x_n)^p.$$

(On obtient le cas précédent pour $x_i = 1$)

Démonstration. C'est pour $n = 2$ la formule du binôme et on étend au cas général par récurrence en appliquant le cas $n = 2$ puis l'hypothèse de récurrence

$$\begin{aligned} (x_1 + \dots + x_n)^p &= \sum_{s_1 + s_2 = p, s_i \geq 0} \frac{p!}{s_1! s_2!} (x_1 + \dots + x_{n-1})^{s_1} x_n^{s_2} \\ &= \sum_{s_1 + s_2 = p, s_i \geq 0} \frac{p!}{s_1! s_2!} \sum_{r_1 + \dots + r_{n-1} = s_1, r_i \geq 0} \frac{s_1!}{r_1! \dots r_{n-1}!} x_1^{r_1} \dots x_{n-1}^{r_{n-1}} x_n^{s_2} \\ &= \sum_{s_1 + s_2 = p, s_i \geq 0} \sum_{r_1 + \dots + r_{n-1} + s_2 = p, r_i \geq 0} \frac{p!}{s_1! r_1! \dots r_{n-1}! s_2!} x_1^{r_1} \dots x_{n-1}^{r_{n-1}} x_n^{s_2}. \end{aligned}$$

Ceci conclut en ré-indiquant $r_n = s_2$. \square

Exemple 12. On veut ranger sur un rayon dix ouvrages comprenant 5 livres de chimie, 3 livres de mathématiques et 2 livres de physique. Le nombre de rangements possibles est celui des arrangements avec répétition d'ordre $(5; 3; 2)$ de 10 objets parmi 3. Ce nombre est $\frac{10!}{5!3!2!} = 2620$.

Exemple 13. Dans une compétition, le nombre de classements où figurent 3 français, 4 anglais et 2 allemands est celui des arrangements avec répétition d'ordre $(3; 4; 2)$ de 9 objets parmi 3. Ce nombre est $\frac{9!}{3!4!2!} = 1260$.

4.6 Combinaison avec répétition (facultatif)

Définition 9. Soit $\Omega = \{a_1, \dots, a_n\}$ un ensemble à n éléments, p un entier et $(r_1; \dots; r_n)$ une suite d'entiers telle que $\sum_{k=1}^n r_k = p$. On appelle *combinaison avec répétition*, de longueur p et d'ordre $(r_1; \dots; r_n)$ tout ensemble de p objets où, pour tout $k \in \llbracket 1, n \rrbracket$, r_k objets sont égaux à a_k . Ce qui compte ici, c'est uniquement le nombre de fois où chaque objet apparaît, abstraction faite de l'ordre d'apparition. Il y a évidemment une seule combinaison avec répétition de ce type.

Le nombre total de toutes les combinaisons avec répétition de p objets parmi n , noté K_n^p , est

$$K_n^p = \binom{n+p-1}{p} = \binom{n+p-1}{n-1}.$$

On peut noter qu'une combinaison avec répétition de p objets parmi n est l'image (avec multiplicité) d'une application $f : \llbracket 1, p \rrbracket \rightarrow \Omega$ telle que $\sum_{k=1}^n \text{Card}(f^{-1}(a_k)) = p$. C'est aussi le nombre de p -tirages avec remise dans une population de taille n où seul compte le nombre d'apparitions de chaque individu. (Pour s'en souvenir et montrer la formule, on comprend qu'une telle combinaison revient à séparer n classes (avec $n-1$ délimiteurs) dans l'ensemble à p éléments, ce qui revient à choisir les $n-1$ délimiteurs dans un ensemble à $n-1+p$ éléments formés des délimiteurs et des éléments de base).

Exemple 14. On peut répartir 10 voix entre 3 candidats de $K_3^{10} = \binom{12}{2} = 12 \cdot 11 / 2 = 66$ façons.

Exemple 15. On peut répartir 8 personnes selon leur groupe sanguin (A, B, AB, O) de $K_4^8 = \binom{11}{3} = 11 \cdot 10 \cdot 9 / 6 = 3 \cdot 5 \cdot 5 = 165$ façons.

Exemple 16. Le développement (multinomial) de $(a+b+c+d)^7$ comprend $K_4^7 = \binom{10}{3} = 10 \cdot 9 \cdot 8 / 6 = 120$ termes.

5 Application : espaces de probabilités finis

On traitera plus en détail au chapitre 3 la modélisation probabiliste sur un ensemble de réalisations possibles dénombrable. On présente juste ici rapidement sur le cas des espaces finis les principales notions utilisées ensuite dans le cours.

L'espace des réalisations Ω d'une expérience aléatoire est l'ensemble des résultats possibles de cette expérience. Cet espace est aussi appelé univers, référentiel, ensemble fondamental ou espace des épreuves. L'ensemble Ω sera fini dans cette section.

Les éléments $\omega \in \Omega$ sont appelés réalisations de l'expérience en question. Ce sont chacun des résultats possibles de l'expérience.

Dans l'étude d'une expérience aléatoire, on s'intéresse à des ensembles de réalisations appelés évènements. Un évènement lié à une expérience aléatoire et à son espace Ω s'identifie donc à une partie A de Ω . Les évènements élémentaires s'identifient aux singletons $\{\omega\}$.

Quand Ω est fini on peut sans problème supposer que l'ensemble des évènements considérés est $\mathcal{T} = \mathcal{P}(\Omega)$.

Exemple 17. Pour un lancer de dé On prend $\Omega = \{1, 2, 3, 4, 5, 6\}$ l'ensemble des face comme espace des réalisations. L'évènement "obtenir une face paire" est représenté par l'ensemble $A = \{2, 4, 6\}$.

Remarque 4. Tout évènement étant une partie de Ω , une correspondance s'établit alors entre le langage naturel des évènements et celui des sous-ensembles. Nous en donnons ici quelques exemples.

- L'espace Ω représente l'évènement certain et l'évènement impossible correspond à l'ensemble vide \emptyset .
- L'évènement contraire à l'évènement A est représenté par son complémentaire A^c .
- Deux évènements A et B sont dits *incompatibles* lorsqu'ils sont disjoints, i.e. lorsque $A \cap B = \emptyset$.

L'objectif des probabilités est d'abord de décrire les probabilités des évènements, d'où l'introduction de la définition suivante qui attribue une probabilité dans $[0, 1]$ à chaque évènement. Le cas le plus simple est celui où tous les éléments de l'espace de réalisation sont équiprobables, on parle de probabilité uniforme. **Dans ce cas très particulier**, la probabilité va être reliée au cardinal des ensembles de façon simple par la formule $P(A) = \text{Card}(A)/\text{Card}(\Omega)$, de sorte qu'on va prendre pour définition d'une probabilité, une fonction qui vérifie le même genre de propriétés que cet exemple. Le cas général correspondra à attribuer des poids différents/chances de réalisations différentes aux différents éléments de l'ensemble.

Définition 10 (Définition d'une probabilité). On appelle *probabilité* sur un ensemble fini Ω toute application $P : \mathcal{P}(\Omega) \rightarrow [0, 1]$ ayant les propriétés suivantes :

1. $P(\Omega) = 1$.
2. Pour toute paire A, B d'évènements incompatibles (c'est-à-dire A, B disjoints) :

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

Remarque 5. Remarquez que $\mathcal{P}(\Omega)$ (avec un \mathcal{P} caligraphique) désigne l'ensemble des parties de Ω , ce qui n'a rien à voir avec la valeur de la fonction P (P droit) sur la partie particulière Ω qui vaut $P(\omega) = 1$

Le couple (Ω, P) est un *espace de probabilité fini*.

Une probabilité sur un ensemble fini $\Omega = \{\omega_1, \dots, \omega_n\}$ est complètement déterminée par la donnée de n nombres positifs $(p_k)_{1 \leq k \leq n}$ tels que $\sum_{k=1}^n p_k = 1$. Le nombre p_k représente alors la probabilité de la réalisation ω_k et, pour tout $A \subset \Omega$, on a

$$P(A) = \sum_{k \in A} p_k.$$

On a les propriétés simples que l'on verra plus en détail par la suite et qui sont similaires aux propriétés du cardinal de la section Dénombrement.

Proposition 19. 1. $P(A^c) = 1 - P(A)$, donc $P(\emptyset) = 0$.

2. Si $A \subset B$ alors $P(B) = P(A) + P(B - A)$. En particulier, si $A \subset B$ alors $P(A) \leq P(B)$.
3. Si $A \cap B = \emptyset$, alors $P(A \cup B) = P(A) + P(B)$. Dans le cas général, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Plus généralement encore,

4.

$$P(\cup_{k=1}^n A_k) = \sum_{k=1}^n (-1)^{k-1} \sum_{1 \leq i_1 < \dots < i_k \leq n} P(A_{i_1} \cap \dots \cap A_{i_k}).$$

(Formule du crible ou de Poincaré, cf TD).

5.1 La probabilité uniforme $\mathcal{U}(n)$ et dénombrement en pratique en probabilité

Lorsque toutes les réalisations sont équiprobables, tous les nombres p_k sont égaux et de somme 1. Il en résulte que, pour tout $k \in \llbracket 1, n \rrbracket$, $p_k = \frac{1}{n}$. La probabilité correspondante P est la probabilité uniforme sur un ensemble de cardinal n , notée $\mathcal{U}(n)$. Cette probabilité modélise toute expérience aléatoire où les réalisations sont tirées ou choisies ou apparaissent "au hasard" parmi n sans qu'aucune ne soit privilégiée par rapport aux autres. Dans ce cas, et uniquement dans ce cas, la probabilité d'un évènement est donnée par la formule

$$P(A) = \frac{\text{Card}(A)}{\text{Card}(\Omega)}.$$

qui se déduit de $P(A) = \sum_{k \in A} \frac{1}{n}$. Cette relation s'exprime en énonçant que la probabilité d'un évènement (pour la probabilité uniforme) est le rapport du nombre de cas favorables au nombre de cas possibles.

Remarque 6. On est souvent amené à réaliser des expériences successivement. Quand on fait une suite de 2 expériences aléatoires, d'espaces d'état Ω_1, Ω_2 , on utilise le produit ensembliste pour décrire le résultat de l'expérience multiple. Les réalisations sont alors les paires de résultats et $\Omega = \Omega_1 \times \Omega_2$. Plus généralement, si l'évènement A se décrit comme la réalisation de deux expériences successives, on peut utiliser la reformulation suivante du principe du Berger :

Proposition 20. (*Principe du dénombrement en probabilité*) *On réalise successivement deux expériences qui peuvent produire respectivement n et m résultats différents. Au total, pour les deux expériences prises ensemble, il existe $n.m$ résultats possibles.*

Démonstration. Formellement A est l'ensemble des résultats possibles et A_1 le résultat de la première expérience. Si on oublie la deuxième expérience en regardant l'application $p : A \rightarrow A_1$ qui associe le résultat de la première expérience à partir du résultat des deux expériences. p est une application et $\{x \in A : p(x) = a\}$ correspond à la réalisation de la deuxième expérience, donc l'hypothèse est que $\text{Card}(\{x \in A : p(x) = a\}) = m$ et $\text{Card}(A_1) = n$. Le principe du Berger (Proposition 6) donne le résultat. \square

On remarquera qu'on aurait pu compter ainsi le nombre d'arrangements, ou de combinaisons, puisque cela remplace le principe du Berger.

Exemple 18. On tire une pièce puis un dé au hasard. Par convention, on dit qu'obtenir P vaut 1, et face vaut 0. Donc $\Omega = \{0, 1\} \times \{1, 2, 3, 4, 5, 6\}$.

On gagne si la pièce et le dé ont la même parité. Au premier tirage, on peut obtenir n'importe quel côté de la pièce $n = 2$ possibles et au deuxième tirage, on a $m = 3$ côté du dé ayant la même parité que la pièce tirée. Au total, on a $\text{Card}(A) = n.m = 6$. Vu $\text{Card}(\Omega) = 2.6 = 12$, on gagne avec probabilité $P(A) = 6/12 = 1/2$.

La description ensembliste est plus compliquée : $A = \{(0, 2), (0, 4), (0, 6), (1, 1), (1, 3), (1, 5)\}$ qui n'est pas exactement un ensemble produit (même si il est en bijection avec un produit $\{0, 1\} \times \{2, 4, 6\}$, par exemple via $(x, y) \mapsto (x, x + y)$)

5.2 Les variables aléatoires et leurs lois

Définition 11. Une variable aléatoire (v.a.) définie sur un espace de probabilité fini (Ω, P) est une application $X : \Omega \rightarrow \mathbb{R}$.

Quand on fait une suite de 2 expériences aléatoires, d'espace d'état Ω_1, Ω_2 , on utilise le produit ensembliste pour décrire le résultat de l'expérience multiple. Les réalisations sont alors les paires de résultats et $\Omega = \Omega_1 \times \Omega_2$. Remarquez que l'on ne fait pas encore d'hypothèse sur la probabilité obtenue sur Ω , on verra au chapitre 4 une relation particulière nommée indépendance (assez naturelle pour que la probabilité uniforme sur Ω donne des variables uniformes indépendantes). Pour $\Omega_i \subset \mathbb{R}$, Les projection $P_i : \Omega \rightarrow \Omega_i$ sont les exemples les plus simples de variables aléatoires dans beaucoup de problèmes.

Exemple 19. Considérons le lancer de deux dés dont l'ensemble des réalisations est $\Omega = \llbracket 1, 6 \rrbracket^2$. A chaque réalisation $\omega = (i, j)$, on peut associer la somme $S(\omega) = i + j$ des points obtenus, le maximum $M(\omega) = \max(i, j)$ des points obtenus, etc. On définit ainsi des variables aléatoires.

Exemple 20. Considérons une succession de tirages au hasard d'une boule dans une urne contenant 4 boules noires et 6 boules rouges. Le rang d'apparition de la première boule rouge est une variable aléatoire.

Le but des probabilités est de décrire des lois générales du hasard, qui ne dépendent pas trop du modèle choisi, à savoir du choix particulier d'espace de probabilité (Ω, P) . Pour décrire, comment, étant donné une variable aléatoire sur un espace particulier, le problème ne dépend pas trop du modèle, on préfère se ramener à une probabilité sur un ensemble de points de \mathbb{R} : la loi de la variable aléatoire. Ainsi, les résultats de la théorie des probabilités ne dépendront que des lois des variables aléatoires considérées.

Rappel 7. Soit $A \subset \mathbb{R}$ et $X : \Omega \rightarrow \mathbb{R}$, on rappelle que l'image réciproque $X^{-1}(A)$ est définie par :

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

On a vu en L1 les relations

$$X^{-1}(A \cup B) = X^{-1}(A) \cup X^{-1}(B),$$

$$X^{-1}(A \cap B) = X^{-1}(A) \cap X^{-1}(B),$$

$$X^{-1}(A^c) = [X^{-1}(A)]^c.$$

En probabilité on note souvent $\{X \in A\}$ au lieu de $X^{-1}(A)$, pour obtenir une notation suggestive.

Définition 12. Soit $X, Y : \Omega \rightarrow \mathbb{R}$ deux v.a sur un ensemble fini Ω . X, Y sont *indépendantes* si pour tout $s \in S = X(\Omega), t \in T = Y(\Omega)$ on a

$$P(X = s, Y = t) = P(X = s)P(Y = t).$$

Exemple 21. Considérons le lancer de deux dés dont l'ensemble des réalisations est $\Omega = \llbracket 1, 6 \rrbracket^2$. A chaque réalisation $\omega = (i, j)$, on peut associer $X_1(\omega) = i, X_2(\omega) = j$, et X_1, X_2 sont indépendantes.

Définition 13. Soit $X_1, \dots, X_n : \Omega \rightarrow \mathbb{R}$ n v.a sur un ensemble fini Ω . X, Y sont *indépendantes* si pour tout $x_i \in X_i(\Omega)$ on a

$$P(X_1 = s_1, \dots, X_n = s_n) = \prod_{i=1}^n P(X_i = s_i).$$

Définition 14. Soit $X : \Omega \rightarrow \mathbb{R}$ une v.a sur un ensemble fini Ω . On appelle loi (ou distribution) de X , la probabilité sur $S = X(\Omega)$, notée P_X , définie par

$$P_X(B) = P(\{X \in B\}), B \in \mathcal{P}(S)$$

On note souvent simplement $P(X \in B)$ pour $P(\{X \in B\})$. On vérifie que P_X ainsi définie est effectivement une probabilité.

En effet, $X^{-1}(\emptyset) = \emptyset$ donc $P_X(\emptyset) = P(\emptyset) = 0$, de même $P_X(S) = P(\Omega) = 1$ et enfin, pour A, B disjoints, $X^{-1}(A), X^{-1}(B)$ le sont aussi et donc

$$P_X(A \cup B) = P(X^{-1}(A) \cup X^{-1}(B)) = P(X^{-1}(A)) + P(X^{-1}(B)) = P_X(A) + P_X(B).$$

Divers paramètres et fonctions permettent de caractériser la loi d'une v.a.d. On commence par la moyenne qui donne une idée de l'ordre de grandeur de la variable en question.

5.3 Moyenne ou Espérance ou Moment d'ordre 1

Définition 15. Pour v.a. X sur une espace de probabilité fini à valeur dans $S = X(\Omega)$, on appelle Moyenne ou Espérance ou Moment d'ordre 1 de X la quantité

$$m = E[X] = \sum_{s \in S} sP(X = s).$$

On verra au chapitre 6 le résultat fondamental suivant (que l'on admet pour l'instant) :

Théorème 21. Soient $X, Y : \Omega \rightarrow \mathbb{R}$ deux v.a.d., sur un espace de de probabilité (fni) alors $E(X + Y) = E(X) + E(Y)$.

Le théorème de transfert permet alors de calculer plus simplement l'espérance de v.a. qui sont des fonctions d'une v.a. donnée.

Théorème 22. Soit $X : \Omega \rightarrow \mathbb{R}$ une v.a.d., $S = X(\Omega)$ et $h : S \rightarrow \mathbb{C}$ une fonction. Alors, la v.a.d. $h(X)$ a pour moyenne :

$$E[h(X)] = \sum_{s \in S} h(s)P(X = s).$$

Démonstration. On a l'union disjointe finie

$$\{h(X) = y\} = \cup_{x \in X(\Omega): h(x)=y} \{X = x\}$$

Donc en utilisant la probabilité des unions disjointes puis en réordonnant la sommation (sommation par paquets finis) :

$$E(h(X)) = \sum_{y \in h(S)} yP(h(X) = y) = \sum_{y \in h(S)} \sum_{x \in X(\Omega): h(x)=y} h(x)P(X = x) = \sum_{x \in S} h(x)P(X = x).$$

□

A partir d'espérances de fonctions, on déduit d'autres grandeurs significatives :

Définition 16. Le *moment d'ordre k* de X est défini par la quantité

$$m_k = E[X^k] = \sum_{s \in S} s^k P(X = s).$$

Définition 17. La *variance* de X est définie par la quantité $v = \sigma^2 = V[X] = E[(X - m)^2] = E[X^2] - m^2$. L'*écart-type* de X est la quantité $\sigma = (V[X])^{1/2}$.

La variable centrée réduite associée à X est $U = \frac{X-m}{\sigma}$. On a $E[U] = 0$ et $V[U] = 1$.

On verra au chapitre 6 le résultat fondamental suivant (que l'on admet pour l'instant) :

Théorème 23. Soient $X, Y : \Omega \rightarrow \mathbb{R}$ deux v.a.d. INDEPENDANTES, sur un espace de probabilité (fini) alors $V(X + Y) = V(X) + V(Y)$.

Des variables aléatoires classiques vont donner des lois finies classiques (pour dire de façon abrégée lois de probabilités classiques sur un espace de probabilité finie).

5.4 Quelques lois usuelles sur un ensemble fini

La loi uniforme $\mathcal{U}(n)$, de paramètre $n \in \mathbb{N}^*$ a été présentée ci-dessus.

Définition 18. La *loi binomiale* $B(n, p)$ de paramètres $(n, p) \in \mathbb{N}^* \times [0, 1]$. C'est la probabilité sur $\llbracket 0, n \rrbracket$ définie par les nombres $p_k = \binom{n}{k} p^k (1-p)^{n-k}$. La loi $B(1, p)$ est aussi notée $B(p)$ et appelée *loi de Bernoulli* de paramètre p .

Remarque 8. C'est la loi du nombre de succès en n essais identiques et indépendants, lorsque la probabilité du succès à chaque essai est p . (cf chapitre 4 section 2.1 pour cette interprétation).

Définition 19. La *loi hypergéométrique* (ou loi des sondages) $H(N, n, p)$, $n \in \llbracket 1, N \rrbracket$, $p = m/N$. C'est la probabilité sur $\llbracket 0, n \rrbracket$ définie par les nombres $p_k = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}}$.

Notons que certains p_k peuvent être nuls.

Remarque 9. C'est la loi du nombre de succès (c'est-à-dire, le nombre de boules d'une catégorie, disons rouge) obtenus en n tirages **sans remise**, lorsque le nombre initial des boules rouge est m et le nombre total de boules dans l'urne est initialement N . En effet on modélise l'urne par l'espace des réalisations $\Omega = P_n(\llbracket 1, N \rrbracket)$ pour les boules représentés par les entiers de 1 à m sont rouges et celles de $m+1$ à N sont, disons, blanches. On met la probabilité uniforme sur Ω . On obtient la probabilité p_k en évaluant la probabilité de l'évènement $A_{k,n}$ correspondant au tirage de k boules rouge soit une partie de $P_k(\llbracket 1, m \rrbracket)$ et $n-k$ blanches soit une partie de $P_{n-k}(\llbracket 1+m, N \rrbracket)$. On obtient $A_{k,n} = P_k(\llbracket 1, m \rrbracket) \times P_{n-k}(\llbracket 1+m, N \rrbracket)$ de cardinal $\binom{m}{k} \binom{N-m}{n-k}$ d'où la formule pour la probabilité

$$p_k = \frac{\text{Card}(A_{k,n})}{\text{Card}(\Omega)} = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}}.$$

Dans un sondage, l'urne est une population, les boules rouges les supporter d'un candidat A et les boules blanches, les supporters d'un autre candidat. On choisit n la taille de l'échantillon représentatif interrogé dans le sondage. On prend donc au hasard ces n individus, d'où le tirage. La loi décrit la probabilité obtenu d'avoir une réponse favorable au candidat A .

Chapitre 2

Statistiques descriptives : Première partie

La statistique vise à analyser efficacement des “ensembles” d’observations (appelés *données*) souvent grand (d’où l’utilisation de résultats asymptotiques).

Le sujet principal de la statistique asymptotique est de prouver des théorèmes du type de la loi des grands nombres ou du théorème central limite (cf. chap 8) et d’en déduire quelle est la meilleure observation ou le meilleur calcul à faire sur les observations pour obtenir efficacement une information cherchée. C’est donc un domaine à l’intersection des probabilités et de l’optimisation. Le sujet de ce cours et du cours de L3 seront plus élémentaires. On se contentera d’introduire quelques problématiques de statistique déduits des résultats classiques que nous verrons (LGN, TCL).

Mais avant cela, le but premier est de présenter les notions de statistiques descriptives, la partie des statistiques qui vise à synthétiser graphiquement ou numériquement un grand nombre de données.

Dans ce chapitre et en TP, nous utiliserons le logiciel *R*. Nous recommandons vivement le livre *Le logiciel R : maîtriser le langage effectuer des analyses statistiques* de Pierre LAFAYE DE MICHEAUX, Rémy DROUILHET et Benoit LIQUET aux éditions Springer. Il est disponible à la BU et en version électronique (achetée par notre BU) :

<http://docelec.univ-lyon1.fr/login?url=http://link.springer.com/book/10.1007/978-2-8178-0115-5>

Nous illustrerons notre propos avec une version simplifiée du jeu de données statistiques sur l’alimentation des personnes âgées tirée du livre et disponible à la page :

<http://www.biostatisticien.eu/springerR/jeuxDonnees4.html>

Au contraire du livre qui présente systématiquement *R*, nous ne ferons qu’illustrer les notions statistiques de base pour rendre notre propos plus concret.

1 Types de variables statistiques et leurs Représentations graphiques

Commençons par un peu de vocabulaire. Une observation statistique est réalisée sur une *population*. C’est un ensemble dont les éléments dits *individus*, ont des caractères que l’on observe. En général, on ne réalisera des observations que sur une partie de la population, *l’échantillon (statistique)* et les caractères observés, dits *variables statistiques* ne seront donc connus que pour les individus de l’échantillon. L’ensemble des individus de l’échantillon et pour chaque individu, la liste des variables statistiques forment les *données statistiques*.

Remarque 10. L’échantillon (statistique) ne doit pas être confondu avec *l’échantillon probabiliste* qui est un vecteur (X_1, \dots, X_n) de variables indépendantes et de même loi. Le nom commun vient du

fait que l'on suppose souvent que les échantillons statistiques sont des réalisations d'échantillons probabilistes qui peuvent servir à leur modélisation.

Un ensemble de données statistiques courant est souvent enregistré dans un tableur ou un fichier texte accompagné d'un codage de l'information recueillie. Par ex, voici le fichier rassemblant notre sélection des résultats recueillis dans une étude sur l'alimentation d'un échantillon de personnes âgées résidant à Bordeaux (Gironde, France), interrogées en 2000 dans le cadre d'une enquête nutritionnelle. L'échantillon est constitué de 226 sujets.

<i>sexe</i>	<i>the</i>	<i>cafe</i>	<i>taille</i>	<i>poids</i>	<i>age</i>	<i>viande</i>	<i>poisson</i>	<i>matgras</i>
<i>F</i>	0	0	151	58	72	4	3	6
<i>F</i>	1	1	162	60	68	5	2	4
<i>F</i>	0	4	162	75	78	3	1	4
<i>F</i>	0	0	154	45	91	0	4	2
<i>F</i>	2	1	154	50	65	5	3	2
<i>F</i>	2	0	159	66	82	4	2	3
<i>F</i>	2	0	160	66	74	3	3	6
<i>H</i>	0	3	166	80	78	5	0	4
...								

fourni avec l'information suivante extraite du livre cité ci-dessus :

Description	Unité ou Codage	Variable
Sexe	F=Femme; H=Homme	sexe
Consommation journalière de thé	Nombre de tasses	the
Consommation journalière de café	Nombre de tasses	cafe
Taille	cm	taille
Poids	kg	poids
Age le jour de l'entretien	Années	age
Consommation de viande	0=Jamais 1=Moins d'une fois par semaine 2=Une fois par semaine 3=2/3 fois par semaine 4=4/6 fois par semaine 5=Tous les jours	viande
Consommation de poisson	Idem	poisson
Matière grasse préférentiellement utilisée pour la cuisson	1=Beurre 2=Margarine 3=Huile d'arachide 4=Huile de tournesol 5=Huile d'olive 6=Mélange d'huile (type Isio4) 7=Huile de colza 8=Graisse de canard ou d'oie	matgras

La première tâche est de distinguer différents types de variables statistiques, qui auront différentes représentations appropriées, on en distingue classiquement 4 principaux :

1. Les *variables qualitatives nominales* comme le sexe ou la consommation de matière grasse dans l'exemple qui prennent valeur dans un ensemble fini sans notion d'ordre naturel sur ses

éléments (type *factor* dans R)

2. Les *variables qualitatives ordinales* comme la consommation de viande ou de poisson dans l'exemple, prennent valeur dans un ensemble fini avec une notion d'ordre naturel (manger 2 à 3 fois de la viande par semaine est plus que jamais) mais sans qu'une valeur numérique précise puisse être attribuée à chaque classe (sans un choix en parti arbitraire). (type *ordered* dans R)
3. Les *variables quantitatives discrètes* comme le nombre de tasses de café ou de thé bues par jour dans l'exemple, prennent valeur dans un ensemble fini ou dénombrable et sont représentables par des v.a. discrètes. (type *numeric* et en général *int*)
4. Les *variables quantitatives continues* comme la taille, l'âge ou le poids dans l'exemple, prennent valeur dans les réels même si ils sont recueillis avec une certaine précision (qui les ramènent à des variables discrètes avec un grand nombre de valeurs possibles). (type *numeric*)

On remarquera qu'on rassemble parfois les résultats obtenus pour des variables qualitatives discrètes par classes qui sont des intervalles et la variable à valeur dans l'ensemble des classes peut alors être traitée comme une variable qualitative ordinale.

Dans R, des types de données permettent de gérer les différents types de variables statistiques. Chaque type a des fonctions prédéfinies différentes adaptées (R est un langage orienté objet, qui définit des classes pour chacun de ces types d'objets).

Dans l'exemple ci-dessus, chaque ligne correspond à un individu différent et les données sont incorporées dans R sous la forme d'un tableau individu×variables dits *data.frame*. Vous verrez en TP les vecteurs et matrices dont toutes les données doivent être du même type par exemple le plus souvent numérique ou entière. Mais dans un *data.frame*, les données peuvent être de types différents, comme le texte (type "*character*") "H" ou "F" spécifiant le sexe dans l'exemple précédent. Par exemple, on recueille les données précédentes par les commandes :

```
> nutriage<-read.delim(file=file.choose()) # <- affecte le data.frame recueilli en lisant le fichier choisi dans la fenêtre qui s'ouvre à la variable nommée nutriage
> attach(nutriage) #cela rend utilisable les variables comme "the" et les autres titres de colonnes etc. vues comme des vecteurs
```

Nous allons passer en revue comment entrer ce type de données dans R, ou comment attribuer un type de données à une variable d'un *data.frame*, puis comment représenter graphiquement ces variables.

1.1 Variable qualitative nominale et représentations de leurs tables de contingence

Une variable qualitative nominale peut se créer sous R à partir d'un vecteur (créé par la fonction *c*) par les commandes :

```
> col<-factor(c("bleu", "bleu", "vert", "bleu", "bleu", "marron", "marron"))
> col
[1] bleu bleu vert bleu bleu marron marron
Levels : bleu marron vert
```

On peut aussi typer les variables de notre exemple ainsi :

```
> sexe<-as.factor(sexe)
> sexe
[1] F F F F F F F F F F F F F F F H F F H F H F H H F H F H F H F H H F F H H
[38] H H H F H H H F H F F F F F H F H H H H F F H F H F H H H F F H H F H
```

```

...
Levels : F H et changer le nom des niveaux :
> levels(sexe) <- c("Femme", "Homme")
> sexe
[1] Femme Femme
[12] Femme Femme Femme Femme Homme Femme Femme Homme Femme Homme
... ou en faisant pareil pour la variable matgras, on peut obtenir une table de contingence des deux
variables marquant l'effectif commun :
> matgras <- as.factor(matgras)
> levels(matgras) <- c("beurre", "margarine", "arachide", "tournesol", "olive", "Melange",
"colza", "canard")
> TCsexeMatgras <- table(sexe, matgras) qui donne à l'affichage (par exemples 32 femmes de
l'échantillon cuisent à l'huile d'arachide) :

```

sexe	matgras							
	beurre	margarine	arachide	tournesol	olive	Melange	colza	canard
Femme	5	17	32	47	20	18	1	1
Homme	10	10	16	21	20	5	0	3

On peut aussi obtenir la table des fréquences avec les marginales ajoutées :

```

> Tfreq <- TCsexeMatgras / sum(TCsexeMatgras); Total <- sum;
TfreqComplete <- addmargins(Tfreq, FUN = Total, quiet = TRUE)

```

sexe	matgras								Total
	beurre	margarine	arachide	tournesol	olive	Melange	colza	canard	
Femme	0.02212	0.07522	0.14159	0.20796	0.08849	0.07964	0.00442	0.00442	0.62389
Homme	0.04424	0.04424	0.07079	0.09292	0.08849	0.022123894	0.00000	0.01327	0.37610
Total	0.06637	0.11946	0.21238	0.30088	0.17699	0.10176	0.00442	0.01769	1.00000

Ce sont les marginales de cette table (obtenues par exemple par `margin.table(Tfreq, 1)`) ou de la table des effectifs qu'on peut alors représenter par exemple par un diagramme circulaire.

Un diagramme circulaire représente avec un angle $2\pi f$ une colonne du tableau de fréquence f .

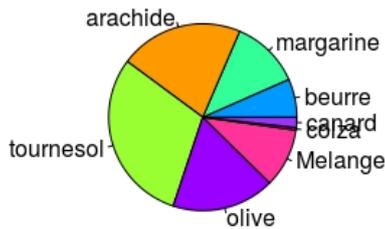
Sur notre exemple, on utilise les commandes (couleurs mises par codage hexadécimal dans un vecteur de texte) :

```

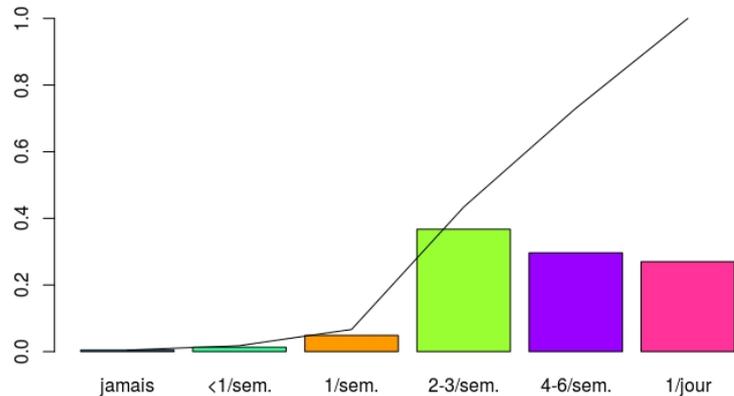
> col <- c("#0099FF", "#33FF99", "#FF9900", "#99FF33", "#9900FF", "#FF3399",
"#FF0099", "#9933FF")
> pie(table(matgras), col = col, main = "Diagramme Circulaire de matgras")
donnant le résultat suivant :

```

Diagramme Circulaire de matgras



Diag. en Tuyau d'orgue de viande



1.2 Variables qualitatives ordinales et diagramme cumulatif

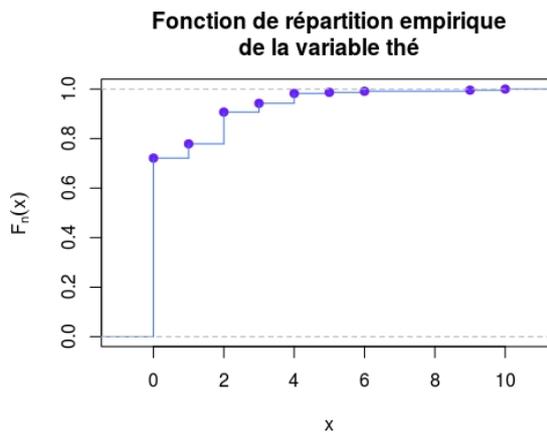
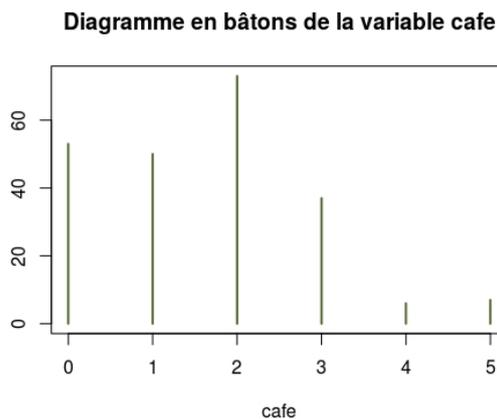
Les variables qualitatives ordinales se définissent de la même manière avec la fonction *ordered()* au lieu de *factor()* ou en typant les variables obtenues par extraction de données (comme ci-dessous). On peut, en plus des fréquences dans la table de contingence (que l'on représente ici par un "diagramme en tuyaux d'orgue" (dit aussi plus simplement diagramme en colonnes) donnant en ordonné la fréquence de chaque classe, qui ressemble mais diffère des histogrammes des variables quantitatives), représenter une fonction de répartition $P(X \leq c)$ pour chacun des niveaux c en faisant la somme cumulée des fréquences (ici interpolé linéairement par convention/choix). Cela donne le résultat ci-dessus avec le code suivant :

```
> viande<-as.ordered(viande);poisson<-as.ordered(poisson)
> niveaux<-c("jamais","<1/sem.,"1/sem.,"2-3/sem.,"4-6/sem.,"1/jour");
  levels(poisson)<-levels(viande)<-niveaux
> tx<-table(viande)
> tx<-tx/sum(tx) # On remplace le tableau des effectifs par celui des fréquences en
divisant par l'effectif total
> abs<-barplot(tx,ylim=c(0,1),col=col,main="Diagramme en Tuyau d'orgue de viande
avec courbe de fréquence cumulée")#la limite des abscisses est [0,1] pour pouvoir tracer
le diagramme cumulé
> points(abs,cumsum(tx),type='l')#cumsum renvoie les sommes partielles de la série
de terme général la suite donnée
```

1.3 Variables quantitatives discrètes et diagramme bâton

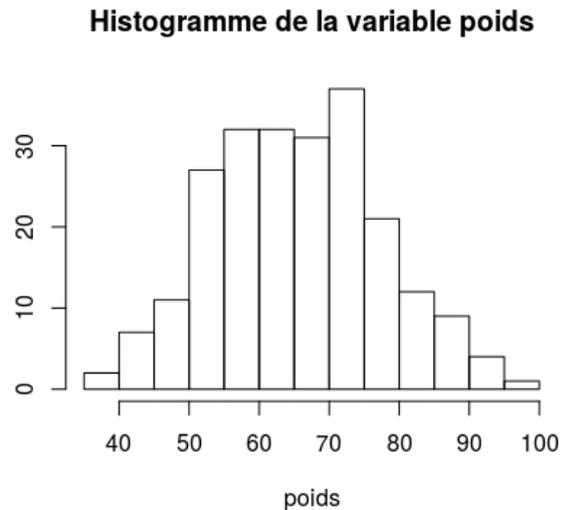
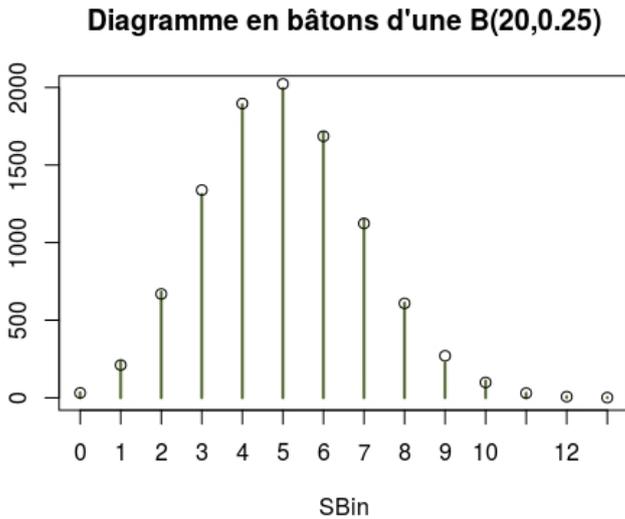
La structuration est facile comme variable entière : `the<-as.integer(the);cafe<-as.integer(cafe)`. On peut alors représenter (ci-dessous à gauche) un diagramme en bâton des effectifs de la population (chaque colonne représente le nombre de personnes dans la table de contingence de la variable) :

```
> plot(table(cafe),ylab="",col="darkolivegreen",main="Diagramme en bâtons de la variable cafe")
```



On peut aussi représenter le graphe de la fonction de répartition empirique (graphe des fréquences cumulées, ou diagramme cumulatif des fréquences, ci-dessus à droite). En partant de la simulation d'une variable discrète, par exemple binomiale $B(20, 0.25)$ (simulée par la fonction *rbinom*, cf TP), on obtient de même le diagramme suivant (les points représentent les effectifs théoriques pour l'échantillon de 10000) :

```
> SBin<-rbinom(10000,20,0.25);SBinTable<-table(SBin);
> plot(SBinTable,col="darkolivegreen",main="Diagramme en bâtons d'une B(20,0.25)")
> points(x=0 :13,y=dbinom(0 :13,20,0.25)*10000,type='p')
```



1.4 Variables quantitatives continues et histogrammes

La structuration est facile comme variable réelle : `poid<-as.double(poid)` ; `taille<-as.double(taille)` ; `age<-as.double(age)`. On pourrait représenter les fonctions de répartition empiriques comme au point précédent. On se concentre ici sur la représentation par histogramme. L'histogramme vise à représenter une approximation rectangulaire de la densité de la variable continue sous-jacente. On découpe l'intervalle $[a, b]$ de valeur des variables (par exemple a le min et b le max des observations) en utilisant une subdivision $a = a_0 < \dots < a_N = b$ donnant N intervalles $I_N = [a_n, a_{n+1}[$ et le dernier $I_{N-1} = [a_{N-1}, a_N]$. On choisit souvent les intervalles de sorte que l'on ait au moins un effectif de 5. On préfère si possible des découpages de pas égaux, mais ce n'est pas toujours adapté.

On calcule alors la hauteur h_n du rectangle de base $[a_n, a_{n+1}[$ pour avoir comme surface la fréquence de la classe. Si N_n est l'effectif de la classe I_N et N_{tot} l'effectif total ($N_{tot} = \sum N_i$) alors

$$h_n = \frac{N_n}{N_{tot}(a_{n+1} - a_n)}.$$

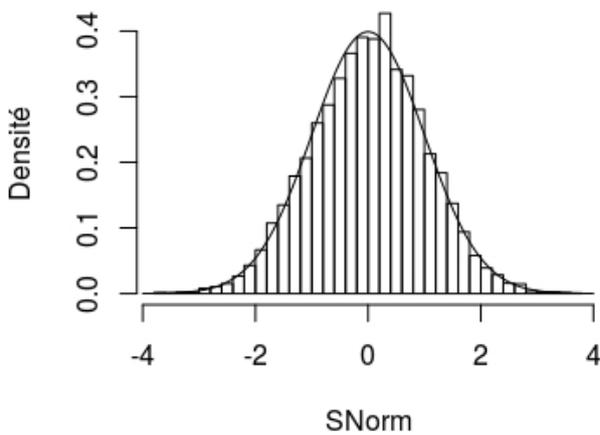
On représente ainsi la densité d'une v.a.c de densité constante sur chaque $[a_n, a_{n+1}[$ et telle que la probabilité de chacun de ces intervalles soit celle de l'échantillon. Si on veut représenter une fonction de répartition associée à cette information par classe, on représente celle de la loi de densité l'histogramme, elle est linéaire par morceau.

Dans *R* sur le cas des simulations de lois puis de la variable âge, on obtient (graphiques ci-dessous) :

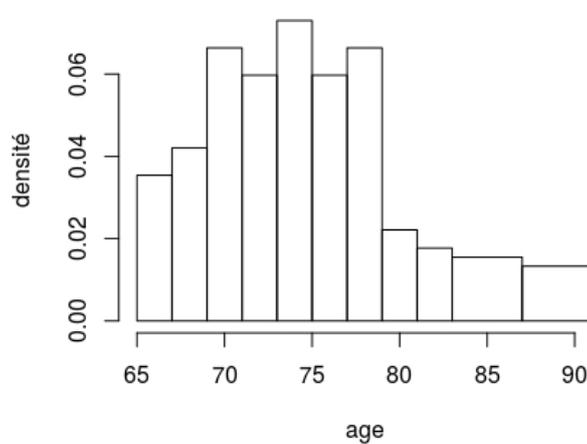
```
> SNorm<-rnorm(10000,mean=0,sd=1)
>hist(SNorm,freq=FALSE,breaks=50,ylab="Densité",main="Histogramme d'une N(0,1)");#
breaks donne ici le nombre de rectangles,
>lines(x=seq(from=-4, to=4, by=0.01),y=dnorm(seq(from=-4, to=4, by=0.01),mean=0,sd=1))
```

```
>TCage<-table(age)
>hist(age,breaks=c(65+2*0 :9,87,91),ylab="densité",right=F,main="Histogramme de la variable age")
```

Histogramme d'une N(0,1)



Histogramme de la variable age



On peut obtenir comme ci-dessus l’histogramme de poids, qui par défaut a des classes de tailles égales. Dans ce cas, la hauteur est proportionnelle à l’effectif et R gradue en effectif plutôt qu’en densité.

```
>TCpoids<-table(poids)
>hist(poids,main="Histogramme de la variable poids",ylab="effectifs")
```

2 Résumés numériques communs d’une variable statistique quantitative

Comme pour les variables aléatoires, on cherche des grandeurs numériques qui résument les observations. Ces résumés sont en général disponibles seulement pour les variables statistiques quantitatives.

2.1 Moyenne et variance empirique

Le plus simple est de considérer l’échantillon des observations (x_1, \dots, x_n) et de considérer une distribution équiprobable pour les observations et de regarder la moyenne ou variance de cette distribution. En somme on considère la variable aléatoire $x : \Omega = \llbracket 1, n \rrbracket \rightarrow \mathbb{R}$ avec $x(n) = x_n$ et Ω est muni de la probabilité uniforme. La loi de x est appelée *distribution empirique* de l’échantillon.

Définition 20. La *moyenne empirique* d’un échantillon $x = x_1, \dots, x_n$ est la grandeur :

$$m_n(x) = \frac{x_1 + \dots + x_n}{n}$$

et la *variance empirique* est la moyenne des carrés des écarts à la moyenne :

$$\sigma_n(x)^2 = \frac{1}{n} \sum_{k=1}^n (x_k - m_n)^2 = \frac{(x_1 - m_n)^2 + \dots + (x_n - m_n)^2}{n}$$

On a la formule alternative

$$\sigma_n^2 + m_n^2 = \frac{1}{n} \sum_{k=1}^n x_k^2.$$

Ces fonctions d'un échantillon peuvent être considérées comme des "estimateurs" de la moyenne et de la variance quand on les considère sur un échantillon probabiliste, c'est à dire que (X_1, \dots, X_n) v.a. i.i.d. (indépendantes et identiquement distribuées). C'est un problème classique de statistique de savoir quelle estimation donne la réponse la plus efficace.

Une contrainte que l'on met souvent est d'obtenir une estimation "non biaisée" c'est à dire redonnant en espérance la bonne valeur. Pour $m_n(X_1, \dots, X_n)$, on obtient

$$E(m_n(X_1, \dots, X_n)) = \frac{E(X_1) + \dots + E(X_n)}{n} = E(X_1)$$

donc la moyenne empirique est non biaisée.

Mais pour la variance on a en utilisant le développement du carré et l'indépendance :

$$\begin{aligned} E(\sigma_n^2(X_1, \dots, X_n)) &= \frac{1}{n} \sum_{k=1}^n E(X_k^2) - E(m_n^2) \\ &= E(X_1^2) - \frac{1}{n^2} \left(\sum_{k=1}^n E(X_k^2) + 2 \sum_{k=1}^n \sum_{l=1}^{k-1} E(X_k)E(X_l) \right) \\ &= \frac{n-1}{n} E(X_1^2) + \frac{1}{n^2} \left(2E(X_1)^2 \sum_{k=0}^{n-1} k \right) \\ &= \frac{n-1}{n} E(X_1^2) + \frac{1}{n^2} \left(2E(X_1)^2 \frac{(n-1)n}{2} \right) = \frac{n-1}{n} Var(X_1) \end{aligned}$$

on obtient donc un estimateur biaisé. On préfère donc la version non biaisée suivante (c'est la déf de la fonction *var* dans R) :

Définition 21. La *variance empirique non-biaisée* d'un échantillon (x_1, \dots, x_n) est la grandeur :

$$var_n(x) = \frac{1}{n-1} \sum_{k=1}^n (x_k - m_n)^2 = \frac{(x_1 - m_n)^2 + \dots + (x_n - m_n)^2}{n-1}$$

2.2 Médiane et quartiles

Les *fractiles* d'ordre pour $\beta \in [0, 1]$ d'une v.a. est la plus petite valeur x_β telle que $P(X \leq x_\beta) \equiv F_X(x_\beta) \geq \beta$:

$$x_\beta = \inf\{x : P(X \leq x) \geq \beta\}.$$

Dans le cas des v.a. continues c'est l'inf. sur les valeurs telles que $F_X(x_\beta) = \beta$ (et on trouve parfois cette définition). Autrement dit, dans ce cadre, comme la fonction de répartition est continue croissante, x_β est bien défini. Pour les variables discrètes et les fractiles, elles ne sont pas forcément définies si on prend l'égalité d'où la définition plus générale choisie ici.

Pour $\beta = 1/4, 1/2, 3/4$, on obtient les *quartiles* de la distribution.

A partir de données statistiques, on a différents choix d'estimateurs de ces quartiles (R propose plusieurs options de type pour la fonction quantile). Le plus simple pour une variable discrète est de remplacer la fonction de répartition par sa variante empirique pour un échantillon $x = (x_1, \dots, x_n)$

$$F_x(t) = \frac{1}{n} \sum_{k=1}^n 1_{\{x_k \leq t\}} = \frac{\text{Card}(\{k: x_k \leq t\})}{n} :$$

$$q_\beta(x) = \inf\{t \geq \min_i x_i : \frac{1}{n} \sum_{k=1}^n 1_{\{x_k \leq t\}} \geq \beta\} = \inf\{t \geq \min_i x_i : F_x(t) \geq \beta\}.$$

On remarque que $q_0(x) = \min_i x_i, q_1(x) = \max_i x_i$. (Pour une variable continue, on peut aussi utiliser la fonction de répartition continue de l'histogramme.)

<code>quantile(age, type = 1)</code>	0%	25%	50%	75%	100%
	65	70	74	78	91

`> quantile(poids, type=1)`

0% 25% 50% 75% 100%

38 57 66 75 96

`> quantile(cafe, type=1)`

0% 25% 50% 75% 100%

0 1 2 2 5

Le 2ème quartile représente une certaine notion de "milieu" de la distribution, mais ce n'est pas la définition la plus commune :

Définition 22. La *médiane* d'un échantillon (x_1, \dots, x_n) est obtenu de deux façons différentes selon la parité de n . On commence par regarder le réarrangement croissant de l'échantillon (x_1^*, \dots, x_n^*) , c'est à dire l'unique permutation $(x_1^*, \dots, x_n^*) = (x_{\sigma(1)}, \dots, x_{\sigma(n)})$ de la suite (x_1, \dots, x_n) telle que $x_1^* \leq x_2^* \leq \dots \leq x_n^*$.

1. Si $n = 2l + 1$, la médiane est l'élément

$$\text{mediane}(x) = x_{l+1}^* = x_{\frac{n+1}{2}}^*.$$

2. Si $n = 2l$, la médiane est l'élément

$$\text{mediane}(x) = \frac{x_l^* + x_{l+1}^*}{2}.$$

Remarque 11. Si $n = 2l + 1$ la médiane est le deuxième quartile, mais si $n = 2l$ le deuxième quartile est x_l^* , c'est à dire la plus petite valeur de l'échantillon en dessous de la médiane. Ce pourrait être une autre définition raisonnable, mais ce n'est pas la plus commune. La fonction `median` de R renvoie la définition ci-dessus. De même, il y a d'autres définitions des quartiles avec des interpolations linéaires des deux éléments les plus proches de notre définition des quartiles.

Exemple 22. Dans l'échantillon $x = (1, 2, 4, 93)$, $F_x(1) = 1/4, F_x(2) = 1/2, F_x(4) = 3/4$ donc 1, 2, 4 sont les 3 quartiles. La médiane est $\frac{2+4}{2} = 3$. La moyenne est $\frac{1+2+4+93}{4} = 25$. On voit que la moyenne prend plus en compte les valeurs extrêmes même si elles sont de faible probabilité.

2.3 Diagrammes à moustaches (ou Diagrammes à boîtes, boxplot)

Un Diagramme à moustache représente les trois quartiles d'une distribution dans une boîte (la barre au milieu donnant la médiane, la barre inférieure le premier quartile, la barre supérieure le deuxième quartile).

Ensuite différentes conventions s'appliquent pour la moustache qui vise à représenter la majeure partie de la distribution. Les ronds à part représente des valeurs extrêmes/aberrantes.

Par défaut dans R ces limites sont placées en dessous et au dessus respectivement, des barres des premiers et troisième quartile respectivement, à une distance de 1,5 fois la distance de la barre du premier au 3ème quartile. Ces barre peuvent être le min ou le max si plus proche de la médiane. Prendre le min et le max (et donc pas de valeurs aberrantes) ou bien un autre fractile sont aussi des choix communs.

Sur la variable café du tableau suivant (nombre de cafés par jours) La médiane et le troisième quartile coïncident à 2, l'écart interquartile est de 1, d'où le haut de la moustache à 3.5, donc les 2 valeurs 4 et 5 sont extrêmes. Le min est 0 et non $1-1.5=-0.5$. Seul l'âge a aussi une valeur extrême, 91, dans les diagrammes de l'âge du poids et de la taille. La commande est `boxplot(list(age,poids,taille),main="Diagramme à moustache de age, poids et taille")`

Diagramme à moustache de age, poids et taille

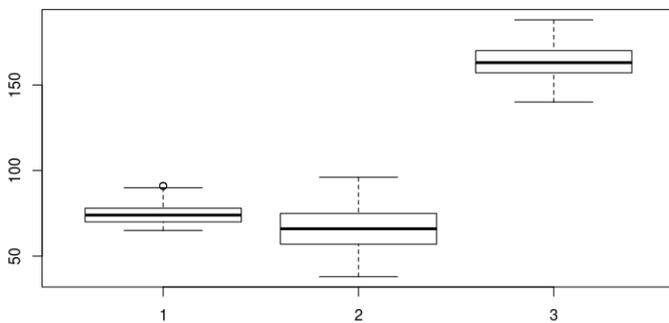
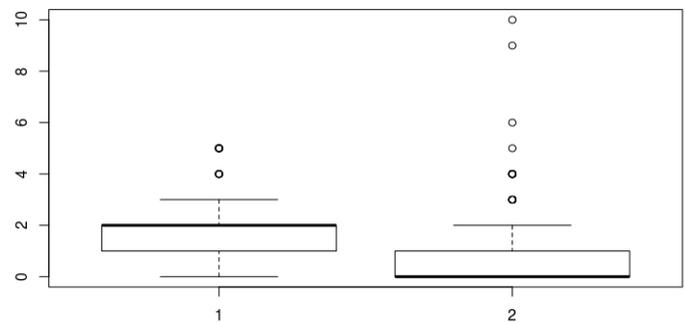


Diagramme à moustache de the et café



Chapitre 3

Familles sommables

Un espace de probabilité discret (disons infini dénombrable au sens où on va bientôt le définir) va associer des nombres, les probabilités aux événements de base $\{\omega_i\}$, correspondant aux éléments ω_i de l'espace des réalisations et en sommant à des événements plus compliqués. Comme ces nombres vont être associés à des ensembles, l'ordre de sommation de ces nombres ne devra pas importer. On va donc étudier une notion de sommation de série où l'ordre de sommation n'importe pas. Le but est donc pour une famille de nombres $(u_i)_{i \in I}$, indicée par un ensemble infini I (le plus souvent dénombrable) de définir la somme :

$$\sum_{i \in I} u_i,$$

en conservant les propriétés de commutativité et d'associativité des sommes finies.

Même dans le cas $I = \mathbb{N}$, le but est d'obtenir une notion de sommation qui ne privilégie pas les sous-ensembles finis $\llbracket 0, n \rrbracket$ comme la notion de somme de série usuelle. On verra que dans ce cas, cette notion de sommation coïncide avec la convergence absolue que vous connaissez déjà.

1 Complément au chapitre 1 : Ensembles infinis

Un ensemble A qui n'est pas fini est dit infini.

1.1 Ensembles infinis dénombrables

Définition 23. Un ensemble infini A est *dénombrable* s'il existe une bijection $f : A \rightarrow \mathbb{N}$. Un ensemble A est *au plus dénombrable* s'il existe une injection $f : A \rightarrow \mathbb{N}$.

Remarque 12. Certains auteurs disent dénombrable pour ce que nous appelons au plus dénombrable et infini dénombrable avec le sens de dénombrable ci-dessus.

On peut représenter les éléments d'un ensemble dénombrable A à l'aide d'une suite infinie en écrivant $A = \{x_n; n \geq 1\}$ (x est l'inverse de la bijection f).

Proposition 24. *Les ensembles au plus dénombrables sont soit finis, soit dénombrables. De plus, pour une partie infinie $P \subset \mathbb{N}$, il existe une bijection strictement croissante et une seule de $\mathbb{N} \rightarrow P$.*

Démonstration. Les ensembles au plus dénombrables sont par définition en bijection avec les parties de \mathbb{N} . Dans le cas infini, il suffit de voir le second point pour obtenir la bijection avec \mathbb{N} . On définit par récurrence la bijection $f : \mathbb{N} \rightarrow P$. Plus précisément, on construit par récurrence sur n une

application strictement croissante $f_n : \llbracket 1, n \rrbracket \rightarrow P$ telle que pour tout $x \in \text{Im}(f_n), y \in P - \text{Im}(f_n), x < y$ et $f_n|_{\llbracket 1, k \rrbracket} = f_k$. Comme P , infini, il est non-vide donc admet un élément $a_0 = \min(P)$ On pose $f_0(0) = a_0$ d'où l'initialisation.

On suppose construit f_n , et on prend $a_{n+1} = \min(P - \text{Im}(f_n))$ qui existe car cette partie est infinie de \mathbb{N} donc non vide (si elle n'était pas infinie, P serait finie comme union finie de parties finies). On pose $f_{n+1}(k) = f_n(k), k \leq n, f_{n+1}(n+1) = a_{n+1}$ de sorte que par l'hyp. de rec. sur $f_n, a_{n+1} > f_n(k), k \leq n$ ce qui donne la stricte croissance de f_{n+1} en combinant avec celle de f_n . Enfin, si $y \in P - \text{Im}(f_{n+1}) \subset P - \text{Im}(f_n)$ on a par hyp de rec $y > f_n(k) k \leq n$ et $y > a_{n+1}$ car c'est le min donc \geq et on a $y \neq a_{n+1}$ par construction. Donc la relation demandée à l'étape suivante est vérifiée.

On obtient f strictement croissante donc injective en rassemblant les valeurs des f_n qui s'accordent ($f(n) = f_n(n) = f_m(n), m \geq n$).

Pour voir que f bijective, par l'absurde, sinon il existe $b \in P - \text{Im}(f)$ mais par stricte croissance d'entiers $f(n) \rightarrow \infty$ donc il existe n minimal tel que $b < f(n) = f_n(n)$ ce qui impose par minimalité $b > f(n-1)$ et contredit $f_n(n) = \text{Min}(P - \text{Im}(f_{n-1}))$ vu $b \in P - \text{Im}(f_{n-1})$.

Pour l'unicité, si g est une autre telle bijection $g^{-1} \circ f$ est une bijection strictement croissante de $\mathbb{N} \rightarrow \mathbb{N}$ ainsi que sa réciproque et le lemme suivant donne donc $g^{-1} \circ f(n) \geq n, f^{-1} \circ g(n) \geq n$ et donc, d'où par croissance de g, f appliquée encore à ces relations : $f = g$. \square

Lemme 25. Une application strictement croissante $f : \mathbb{N} \rightarrow \mathbb{N}$ (resp. $f : \llbracket 0, n \rrbracket \rightarrow \mathbb{N}$) vérifie $f(p) \geq p$ pour tout p dans son domaine.

Démonstration. Il suffit de voir le deuxième cas (en restreignant aux segments initiaux), on le montre par récurrence sur n . Si $n = 0$, $f(0) \in \mathbb{N}$ donc c'est évident. En supposant l'hypothèse vraie au rang n , on considère $f : \llbracket 0, n+1 \rrbracket \rightarrow \mathbb{N}$, la restriction à $\llbracket 0, n \rrbracket$ vérifie l'hypothèse de récurrence, donc $f(p) \geq p$ pour $p \leq n$ et $f(n+1) > f(n) \geq n$ mais dans \mathbb{N} cela implique $f(n+1) \geq n+1$ et conclut l'étape d'induction. \square

Proposition 26. Un ensemble P est au plus dénombrable si et seulement si il existe un surjection $f : \mathbb{N} \rightarrow P$.

Démonstration. Pour l'implication directe, si P est dénombrable, la bijection de la définition convient, si P est fini, en bijection avec $\llbracket 0, n-1 \rrbracket$ alors le reste modulo n donne la surjection $\mathbb{N} \rightarrow \llbracket 0, n-1 \rrbracket$ qui composée à la bijection donne la surjection cherchée. Réciproquement, l'ensemble $f^{-1}(p), p \in P$ est une partie de \mathbb{N} qui a un plus petit élément $a_p : a : P \rightarrow \mathbb{N}$ est l'injection cherchée. \square

On va obtenir des exemples d'ensembles dénombrables les plus courants. Pour cela on a besoin de quelques méthodes de constructions.

Lemme 27. 1. La réunion d'une suite $(X_n)_{n \geq 0}$ d'ensembles finis 2 à 2 disjoints est au plus dénombrable.

2. Un ensemble X est au plus dénombrable si et seulement si il admet une suite exhaustive de parties finies, c'est à dire une suite croissante de parties finies dont l'union est X .

3. Le produit cartésien d'un nombre fini d'ensembles au plus dénombrables est au plus dénombrable.

Démonstration. 1. Soit $a_n = \text{Card}(X_n)$ et $A_n = \sum_{k=0}^n a_k$ ($A_{-1} = 0$). On a des bijections $h_n : \llbracket A_{n-1} + 1, A_n \rrbracket \rightarrow \llbracket 1, a_n \rrbracket \rightarrow X_n$ qui induisent une application $h : \mathbb{N}^* \rightarrow \cup_n X_n$ dès qu'un nombre infini de X_i n'est pas vide, ou $h : \llbracket 1, A_p \rrbracket \rightarrow \cup_n X_n$ qui est par construction surjective. L'injectivité des h_n et le fait que les X_n sont disjoints donne l'injectivité de h . 2. Si X est fini on prend la suite

constante, sinon, pour une bijection $h : \mathbb{N} \rightarrow X$ on prend $X_n = h(\llbracket 0, n \rrbracket)$ comme suite croissante cherchée. Réciproquement, la suite croissante X_n donne une suite disjointe $X_0, X_{n+1} - X_n$ de parties finies, donc 1 donne que l'union est au plus dénombrable.

3. Une récurrence triviale ramène au cas du produit de 2 ensembles A, B . Soit $h : \mathbb{N} \rightarrow A$, $g : \mathbb{N} \rightarrow B$ des surjections données par la proposition 26. $f = h \times g : \mathbb{N}^2 \rightarrow A \times B$ est une surjection qui ramène au cas \mathbb{N}^2 qui admet pour suite exhaustive d'ensembles finis $\llbracket 0, n \rrbracket^2$. \square

Proposition 28. *Les ensembles $\mathbb{N}^k, k \in \mathbb{N}^*$; \mathbb{Z} et \mathbb{Q} sont infinis dénombrables.*

Démonstration. On a vu le cas du produit \mathbb{N}^k au lemme précédent. $\llbracket -n, n \rrbracket$ est une suite exhaustive d'ensemble fini pour \mathbb{Z} qui est donc au plus dénombrable par la proposition précédente, il est infini car il contient \mathbb{N} . Enfin $(p, q) \mapsto p/q$ est une surjection de $\mathbb{Z} \times \mathbb{N}^* \rightarrow \mathbb{Q}$, donc par la proposition 26 \mathbb{Q} est au plus dénombrable, et infini car contient \mathbb{N} . \square

Enfin, on améliore le lemme précédent.

Proposition 29. *Une réunion au plus dénombrable d'ensembles au plus dénombrables est au plus dénombrable.*

Démonstration. Soit $(X_n)_{n \geq 0}$ une suite d'ensembles dénombrables (si la suite est finie, on peut la prolonger en une suite infinie. Petite subtilité, passé d'une famille dénombrable à une suite d'ensemble n'est pas complètement anodin et utilise l'axiome du choix dénombrable). Soit $f_n : \mathbb{N} \rightarrow X_n$ une surjection donnée par la proposition 26. On pose $f : \mathbb{N}^2 \rightarrow \bigcup_{n \in \mathbb{N}} X_n$ défini par $f(n, p) = f_n(p)$ et en composant avec une surjection $\mathbb{N} \rightarrow \mathbb{N}^2$, on obtient le résultat par la réciproque dans la proposition juste citée. \square

Les ensembles au plus dénombrables serviront de base aux probabilités discrètes.

1.2 Ensembles infinis non dénombrables

Les ensembles qui n'appartiennent pas aux catégories précédentes (finis ou infinis dénombrables) sont dits infinis non dénombrables. On va voir que par exemple, \mathbb{R} et \mathbb{C} , $[a, b]$, $a < b$ sont infinis non dénombrables.

Le résultat clef est toujours un argument diagonal :

Lemme 30. *(Théorème de Cantor) Il n'existe pas de surjection $h : E \rightarrow P(E)$ entre un ensemble E et l'ensemble de ses parties.*

Démonstration. En effet une application $h : E \rightarrow P(E)$ permet de considérer l'ensemble $A = \{x \in E : x \notin h(x)\}$. Il n'existe pas de y tel que $h(y) = A$ car par l'absurde, si il existait, soit $y \in A$ et alors $y \notin h(y) = A$ une contradiction, soit $y \notin A$ et alors $y \in h(y) = A$ encore une contradiction. \square

Remarque 13. En conséquence de ce lemme et de la proposition 26, $P(\mathbb{N})$ n'est pas dénombrable (il est infini à cause de l'injection $x \mapsto \{x\}$ défini sur \mathbb{N}), car sinon on aurait une surjection de $\mathbb{N} \rightarrow P(\mathbb{N})$. En conséquence $\{0, 1\}^{\mathbb{N}}$, en bijection avec la fonction indicatrice n'est pas non-plus dénombrable.

Théorème 31. *$[0, 1]$ et \mathbb{R} ne sont pas dénombrables.*

En conséquence un intervalle quelconque $[a, b]$, en bijection avec $[0, 1]$ ne l'est pas non plus. et un intervalle quelconque contenant au moins deux points (qui contient donc aussi un $[a, b]$) est aussi non-dénombrable.

Démonstration. On construit une injection $\varphi : \{0, 1\}^{\mathbb{N}} \rightarrow [0, 1]$ (le cas \mathbb{R} s'en déduit. (l'image de cette injection va être l'ensemble triadique de Cantor). On fixe $a = (a_n) \in \{0, 1\}^{\mathbb{N}}$ on définit une suite de segments emboîtés, on pose $J_0 = [0, 1]$ et si $J_n = [x_n, y_n]$ alors on découpe l'intervalle en trois en posant $u_n = (2x_n + y_n)/3$ et $v_n = (x_n + 2y_n)/3$. Si $a_n = 0$, on pose $J_{n+1} = [x_n, u_n]$, et si $a_n = 1$, on pose $J_{n+1} = [v_n, y_n]$. On obtient par construction une suite de segments emboîtés, x_n, y_n sont des suites adjacentes et $y_n - x_n \leq 1/3^n$ (récurrence facile) donc l'intersection est un singleton $\bigcap_n J_n = \{\varphi(a)\}$.

Pour voir que φ est injective on note que si $a \neq a'$ sont deux suites et n le premier indice avec $a_n \neq a'_n$, alors $J_n \cap J'_n = \emptyset$ et les images sont donc distinctes. \square

Remarque 14. L'ensemble triadique de Cantor a plein de propriétés intéressantes. Topologiquement il est fermé, totalement disconnecté (les composantes connexes sont les singletons). Il est de longueur nulle (car inclus dans l'union sur tous les cas possibles des J_n dont la longueur perd un facteur $2/3$ à chaque n). Le sens de cette longueur sera vu en L3 (c'est la mesure de Lebesgue). Il est en fait fractal de dimension de Hausdorff $\ln(2)/\ln(3) < 1$ (ce qui réexplique la longueur nulle, mais c'est un sujet beaucoup plus avancé).

Exemple 23. L'ensemble des nombres irrationnels $\mathbb{R} - \mathbb{Q}$ est non-dénombrable, car sinon son union avec \mathbb{Q} à savoir \mathbb{R} serait dénombrable, ce qui n'est pas le cas.

2 Familles sommables à termes positifs

Définition 24. Une famille $(a_i)_{i \in I}$ de nombres réels positifs est dite sommable si

$$\sup \left\{ \sum_{j \in J} a_j : J \subset I, \text{ fini} \right\} < \infty$$

et alors on note

$$\sum_{i \in I} a_i = \sup \left\{ \sum_{j \in J} a_j : J \subset I, \text{ fini} \right\}.$$

Tout d'abord, le résultat simple suivant ramène au cas I dénombrable, ce que l'on supposera par la suite :

Lemme 32. Si $(a_i)_{i \in I}$ est une famille sommable, alors le support $I_0 = \{i \in I : a_i \neq 0\}$ est au plus dénombrable.

Démonstration. En effet si $S = \sum_{i \in I} a_i$ et si $I_n = \{i \in I : a_i \geq S/n\}$, alors $I_0 = \bigcup_{n \geq 1} I_n$ est au plus dénombrable comme union d'une suite d'ensembles finis car $\text{Card}(I_n) \leq n$. En effet, si $j \in I_n$, $a_j \geq S/n$ donc si $J \subset I_n$ fini $S \geq \sum_{j \in J_n} a_j \geq S \text{Card}(J)/n$ donc $\text{Card}(J) \leq n$ et donc $\text{Card}(I_n) \leq n$. \square

On résume les propriétés générales dans l'énoncé suivant :

Proposition 33. 1. (critère des suites exhaustives) Si $(J_n)_{n \geq 0}$ est une suite exhaustive de parties finies de I , alors la famille $(a_i)_{i \in I}$ est sommable si et seulement si la suite $(\sum_{i \in J_n} a_i)_{n \geq 0}$ est bornée et alors on a

$$\sum_{i \in I} a_i = \sup_{n \in \mathbb{N}} \sum_{i \in J_n} a_i = \lim_{n \rightarrow \infty} \sum_{i \in J_n} a_i.$$

2. (lemme de domination) Si $a_i \leq b_i$ pour tout i et (b_i) sommable, alors $(a_i)_{i \in I}$ est sommable et alors $\sum_{i \in I} a_i \leq \sum_{i \in I} b_i$.
3. (lemme de permutation) Si $(a_i)_{i \in I}$ est sommable et $\sigma : I \rightarrow I$ est une bijection, alors $(a_{\sigma(i)})_{i \in I}$ est sommable de même somme.

Démonstration. 1/ La famille $\sum_{i \in J_n} a_i$ étant inclus dans la famille des sommes finies, il est clair qu'elle est majorée si la famille est sommable (et on a en passant au sup la partie \geq de l'égalité énoncée). Mais réciproquement toute famille finie est inclus dans un certain J_n , par définition d'une suite exhaustive, d'où la borne inverse et la réciproque.

2/ Il suffit de borner les sommes partielles finies $\sum_{i \in J} a_i \leq \sum_{i \in J} b_i$ et passer au sup.

3/ Pour tout J fini, $\sigma(J)$ est fini donc $\sum_{i \in J} a_{\sigma(i)} = \sum_{i \in \sigma(J)} a_i \leq \sum_{i \in I} a_i$. D'où la sommabilité et la première inégalité en passant au sup. En considérant la bijection réciproque σ^{-1} on obtient de même l'autre inégalité. \square

Le dernier résultat généralise la commutativité des sommes. On conclut avec les deux résultats importants, le premier généralise l'associativité des sommes finies. On rappelle qu'une partition $(I_\lambda)_{\lambda \in \Lambda}$ de I est une famille d'ensembles 2 à 2 disjoints d'union égale à I .

Théorème 34. (de sommation par paquets) Soit $(I_\lambda)_{\lambda \in \Lambda}$ une partition de I . Une famille $(a_i)_{i \in I}$ est sommable si et seulement si on a à la fois les deux propriétés suivantes :

1. pour chaque $\lambda \in \Lambda$, $(a_i)_{i \in I_\lambda}$ est sommable, disons de somme σ_λ
2. et $(\sigma_\lambda)_{\lambda \in \Lambda}$ est sommable.

De plus, on a l'égalité :

$$\sum_{i \in I} a_i = \sum_{\lambda \in \Lambda} \sigma_\lambda \equiv \sum_{\lambda \in \Lambda} \left(\sum_{i \in I_\lambda} a_i \right).$$

Démonstration. Commençons par la condition nécessaire. Si $(a_i)_{i \in I}$ est sommable alors les sommes finies d'une sous famille $(a_i)_{i \in I_\lambda}$ sont bornées par les sommes de la famille totale donc on a la première condition de sommabilité et $\sigma_\lambda \leq \sum_{i \in I} a_i$. Plus si on a des sous ensembles finis $J_1 \subset I_{\lambda_1}, \dots, J_n \subset I_{\lambda_n}$ pour des λ_j distincts, ils sont disjoints et leur union $J = \cup_{k=1}^n J_k$ est un sous-ensemble fini de I donc

$$\sum_{k=1}^n \sum_{i \in J_k} a_i = \sum_{i \in J} a_i \leq \sum_{i \in I} a_i$$

Donc en passant successivement au sup sur les J_k fini, on obtient :

$$\sum_{k=1}^n \sigma_{\lambda_k} \leq \sum_{i \in I} a_i.$$

Donc la famille $(\sigma_\lambda)_{\lambda \in \Lambda}$ est sommable et on obtient la première inégalité \geq en passant au sup.

Réciproquement, pour tout J partie finie de I on définit $J_\lambda = J \cap I_\lambda$ et on obtient un nombre fini de λ tel que $J = \cup_{k=1}^n J_{\lambda_k}$. On déduit

$$\sum_{i \in J} a_i = \sum_{k=1}^n \sum_{i \in J_k} a_i \leq \sum_{k=1}^n \sigma_{\lambda_k} \leq \sum_{\lambda \in \Lambda} \sigma_\lambda.$$

D'où la bornitude sur J qui donne la sommabilité, et l'autre inégalité en passant au sup. \square

Un cas particulier est la "version famille sommable" du théorème de Fubini (qui se généralise à un théorème d'intégration). Le cas positif est nommé théorème de Fubini-Tonelli. Il correspond à la décomposition

$$I \times J = \cup_{i \in I} \{i\} \times J = \cup_{j \in J} I \times \{j\}.$$

Il donne un résultat d'interversion des sommes.

Théorème 35. (de Fubini-Tonelli) Une famille double $(a_{i,j})_{i \in I, j \in J}$ à termes positifs est sommable si et seulement si on a l'une des deux propriétés équivalentes suivantes :

1. pour tout $i \in I$, $(a_{i,j})_{j \in J}$ est sommable et la famille des sommes $(\sum_{j \in J} a_{i,j})_{i \in I}$ est sommable

2. pour tout $j \in J$, $(a_{i,j})_{i \in I}$ est sommable et la famille des sommes $(\sum_{i \in I} a_{i,j})_{j \in J}$ est sommable

Alors, on a l'égalité :

$$\sum_{(i,j) \in I \times J} a_{i,j} = \sum_{i \in I} \left(\sum_{j \in J} a_{i,j} \right) = \sum_{j \in J} \left(\sum_{i \in I} a_{i,j} \right).$$

Démonstration. C'est une application directe du résultat de sommation par paquets avec les partitions ci-dessus. \square

3 Familles sommables à termes scalaires

Comme pour les séries, on se ramène au cas à valeur positif en prenant le module. On pourrait traiter de façon semblable le cas à valeurs vectorielles (par exemple dans \mathbb{R}^n ou dans un e.v.n. où toute suite de Cauchy converge, un e.v.n dit complet) en prenant la norme à la place du module. On note $\mathbb{K} = \mathbb{R}$ ou $\mathbb{K} = \mathbb{C}$ le corps de référence.

Définition 25. Une famille $(z_i)_{i \in I}$ de nombres complexes ou réels est dite sommable si la famille $(|z_i|)_{i \in I}$ est sommable. On note $\ell^1(I, \mathbb{K})$ l'ensemble des familles sommables d'éléments de \mathbb{K} indexées par I .

On note

$$\|z\|_1 = \sum_{i \in I} |z_i|$$

et on va voir que cela définit une norme sur l'espace des familles sommables.

Lemme 36. $(\ell^1(I, \mathbb{K}), \|\cdot\|_1)$ est un espace vectoriel normé.

Démonstration. On voit que c'est un sous-espace vectoriel de l'ensemble des fonctions \mathbb{K}^I . D'abord, la famille nulle est sommable et de plus si $\lambda, \mu \in \mathbb{K}$, $(a_i), (b_i)$ des familles sommables, pour J fini, on a par l'inégalité triangulaire (des nombres) :

$$\sum_{i \in J} |\lambda a_i + \mu b_i| \leq \sum_{i \in J} |\lambda| |a_i| + |\mu| |b_i| = |\lambda| \sum_{i \in J} |a_i| + |\mu| \sum_{i \in J} |b_i| \leq |\lambda| \|a\|_1 + |\mu| \|b\|_1$$

donc comme la valeur est bornée, on obtient, la sommabilité de la famille $(\lambda a_i + \mu b_i)$, donc $\ell^1(I, \mathbb{K})$ est stable par combinaison linéaire et est donc un sous-espace vectoriel de \mathbb{K}^I , puisqu'il contient aussi la famille nulle (0).

De plus en passant au sup sur J on obtient $\|\lambda a + \mu b\|_1 \leq |\lambda| \|a\|_1 + |\mu| \|b\|_1$. Donc $\|\cdot\|_1$ vérifie l'inégalité triangulaire (cas $\lambda = \mu = 1$). De plus $\|\cdot\|_1$ est positif. Comme $|a_i| \leq \|a\|_1$, $a_i = 0$ si $\|a\|_1 = 0$, pour tout i donc $a = 0$ ce qui donne l'axiome de séparation. Enfin $\sum_{i \in J} |\lambda a_i| = |\lambda| \sum_{i \in J} |a_i|$ donc en passant au sup : $|\lambda| \|a\|_1 = \|\lambda a\|_1$ (d'où l'homogénéité). \square

Comme d'habitude pour définir l'intégrale (ici on va définir de même la somme), on sépare les parties positives, négatives des parties réelles et imaginaires, pour définir la somme. On note donc $(a_i)_+ = \max(a_i, 0)$, $(a_i)_- = \max(-a_i, 0)$ de sorte que $z_j = (\operatorname{Re} z_j)_+ - (\operatorname{Re} z_j)_- + i(\operatorname{Im} z_j)_+ - i(\operatorname{Im} z_j)_-$

Comme $(\operatorname{Re} z_j)_+ + (\operatorname{Re} z_j)_-, (\operatorname{Im} z_j)_+ + (\operatorname{Im} z_j)_- \leq |z_j|$ on déduit que si (z_j) est sommable, alors $((\operatorname{Re} z_j)_+), ((\operatorname{Re} z_j)_-), ((\operatorname{Im} z_j)_+), ((\operatorname{Im} z_j)_-)$ le sont aussi par domination.

Définition 26. La somme d'une famille sommable $(z_i)_{i \in I}$ est la valeur :

$$\sum_{j \in I} z_j := \sum_{j \in I} (\operatorname{Re} z_j)_+ - \sum_{j \in I} (\operatorname{Re} z_j)_- + i \sum_{j \in I} (\operatorname{Im} z_j)_+ - i \sum_{j \in I} (\operatorname{Im} z_j)_-$$

Exercice 1. Vérifier que la somme d'une famille sommable est une application linéaire. (indication : considérer une suite exhaustive de parties finies pour se ramener au cas des sommes finies).

On a le résultat qui résume les propriétés élémentaires :

Proposition 37. 1. Une famille $(z_i)_{i \in I}$ est sommable si et seulement si $(\operatorname{Re} z_i)_{i \in I}$ et $(\operatorname{Im} z_i)_{i \in I}$ sont sommables.

2. $(z_i)_{i \in I}$ est sommable si et seulement si $(\overline{z_i})_{i \in I}$ est sommable et on a :

$$\overline{\sum_{j \in I} z_j} = \sum_{j \in I} \overline{z_j},$$

3. Pour $(z_i)_{i \in I}$ sommable, on a l'inégalité triangulaire généralisée :

$$\left| \sum_{j \in I} z_j \right| \leq \sum_{j \in I} |z_j|.$$

4. (lemme de permutation) Si $(z_i)_{i \in I}$ est sommable et $\sigma : I \rightarrow I$ est une bijection, alors $(z_{\sigma(i)})_{i \in I}$ est sommable de même somme. En particulier, si $\sum a_n$ est une série absolument convergente et σ une permutation de \mathbb{N} alors $\sum a_{\sigma(n)}$ est absolument convergente de même somme.

Démonstration. 1/ Les bornes $|\operatorname{Re} z_i| \leq |z_i|$ et $|\operatorname{Im} z_i| \leq |z_i|$ donnent la condition nécessaire par domination. Réciproquement $|z_i| = \sqrt{|\operatorname{Re} z_i|^2 + |\operatorname{Im} z_i|^2} \leq |\operatorname{Re} z_i| + |\operatorname{Im} z_i|$ et comme ℓ^1 est un e.v, on a vu que l'hypothèse implique $(|\operatorname{Re} z_i| + |\operatorname{Im} z_i|)_{i \in I}$ sommable d'où le résultat à nouveau par domination.

2/ l'équivalence est évidente en utilisant 2 fois le 1. L'égalité vient directement de la définition.

3/ On fixe une suite exhaustive J_n de I . D'après le critère des suites exhaustives pour les quatre séries à termes positives intervenant dans la somme, $\sum_{j \in I} z_j = \lim_{n \rightarrow \infty} \sum_{j \in J_n} z_j$, $\sum_{j \in I} |z_j| = \lim_{n \rightarrow \infty} \sum_{j \in J_n} |z_j|$ donc par l'inégalité triangulaire pour les sommes finies (et continuité du module)

$$\left| \sum_{j \in I} z_j \right| = \left| \lim_{n \rightarrow \infty} \sum_{j \in J_n} z_j \right| = \lim_{n \rightarrow \infty} \left| \sum_{j \in J_n} z_j \right| \leq \lim_{n \rightarrow \infty} \sum_{j \in J_n} |z_j| = \sum_{j \in I} |z_j|.$$

4/ Tout vient du cas positif, soit par la définition de sommabilité soit par la définition de la somme en terme de somme de familles à termes positifs. Le cas particulier vient du fait que si la famille est indicée par \mathbb{N} , le critère des suites exhaustives (appliqué à la suite $\llbracket 0, n \rrbracket$) implique qu'être sommable équivaut à être absolument convergente. \square

Remarque 15. Une série telle que $\sum a_n$ telle que pour tout σ permutation de \mathbb{N} on ait $\sum a_{\sigma(n)}$ convergent est dite inconditionnellement convergente. Un résultat classique qu'on trouve par exemple dans Bourbaki Topologie Générale III.44 dit qu'une série numérique inconditionnellement convergente est absolument convergente. Il n'y a donc pas d'extension possible du dernier énoncé.

On finit avec les résultats de sommation par paquets et de Fubini. Dans les deux cas, on n'a plus d'équivalence comme dans le cas à terme positif. On utilise alors souvent/toujours le cas à terme positif pour montrer la sommabilité nécessaire à appliquer le cas avec signe/complexe.

Théorème 38. (de sommation par paquets) Soit $(I_\lambda)_{\lambda \in \Lambda}$ une partition de I . Si une famille $(z_i)_{i \in I}$ est sommable alors on a les deux propriétés suivantes :

1. pour chaque $\lambda \in \Lambda$, $(z_i)_{i \in I_\lambda}$ est sommable, disons de somme σ_λ
2. et $(\sigma_\lambda)_{\lambda \in \Lambda}$ est sommable.

De plus, on a l'égalité :

$$\sum_{i \in I} z_i = \sum_{\lambda \in \Lambda} \sigma_\lambda \equiv \sum_{\lambda \in \Lambda} \left(\sum_{i \in I_\lambda} z_i \right).$$

Démonstration. Comme $(|z_i|)_{i \in I}$, la sommabilité de $(|z_i|)_{i \in I_\lambda}$ vient du cas positif. De plus, par l'inégalité triangulaire des familles sommables (proposition 37), $|\sum_{i \in I_\lambda} z_i| \leq \sum_{i \in I_\lambda} |z_i|$ et le théorème de sommation par paquets assure la sommabilité du membre de droite, donc par comparaison, celle de $(\sigma_\lambda)_{\lambda \in \Lambda}$ comme voulu. L'égalité vient du cas positif appliqué aux parties positives et négatives des parties réelle et imaginaire. \square

En appliquant la sommation par paquets à la même partition que dans le cas positif, on obtient :

Théorème 39. (de Fubini) Si une famille double $(z_{i,j})_{i \in I, j \in J}$ est sommable alors on a les deux propriétés suivantes :

1. pour tout $i \in I$, $(z_{i,j})_{j \in J}$ est sommable et la famille des sommes $(\sum_{j \in J} z_{i,j})_{i \in I}$ est sommable
2. pour tout $j \in J$, $(z_{i,j})_{i \in I}$ est sommable et la famille des sommes $(\sum_{i \in I} z_{i,j})_{j \in J}$ est sommable

De plus, on a l'égalité :

$$\sum_{(i,j) \in I \times J} z_{i,j} = \sum_{i \in I} \left(\sum_{j \in J} z_{i,j} \right) = \sum_{j \in J} \left(\sum_{i \in I} z_{i,j} \right).$$

4 Familles de carrés sommables et espaces $L^p(\Omega, P)$

Il est souvent utile de considérer des produits de familles sommables ou de savoir quand un produit de suites est sommable quand les suites de départ ne le sont pas.

On commence par introduire une notion qui interviendra peu dans la suite de ce cours mais plus comme exemple dans le cours d'intégration de L3. Elle introduit un e.v.n crucial qui généralise l'espace euclidien \mathbb{R}^n , c'est l'espace de Hilbert $\ell^2(\mathbb{N})$. Ensuite, on introduit des variantes probabilistes qui interviendront pour parler d'invariants de bases des lois de probabilités (discrètes) comme l'espérance ou la variance.

4.1 Familles de carré sommable (facultatif)

Définition 27. Une famille $(z_i)_{i \in I}$ de nombres complexes ou réels est dite de carré sommable si la famille $(|z_i|^2)_{i \in I}$ est sommable. On note $\ell^2(I, \mathbb{K})$ l'ensemble des familles d'éléments de \mathbb{K} de carré sommable.

On note

$$\|z\|_2 = \left(\sum_{i \in I} |z_i|^2 \right)^{1/2}.$$

Le théorème suivant résume les propriétés essentielles :

Théorème 40.

1. Si $(x_i)_{i \in I}$ et $(y_i)_{i \in I}$ sont de carré sommable alors $(x_i y_i)_{i \in I}$ est sommable.
2. $\ell^2(I, \mathbb{K})$ est un sous-espace vectoriel de \mathbb{K}^I .
3. On a l'inégalité de Cauchy-Schwarz :

$$\left| \sum_{i \in I} x_i y_i \right|^2 \leq \left(\sum_{i \in I} |x_i|^2 \right) \left(\sum_{i \in I} |y_i|^2 \right).$$

Démonstration. 1/ L'inégalité arithmetico géométrique $|x_i y_i| \leq (|x_i|^2 + |y_i|^2)/2$ donne la sommabilité par domination.

2/ La formule $|x_i + \lambda y_i|^2 = |x_i|^2 + |\lambda|^2 |y_i|^2 + 2 \operatorname{Re}(\overline{x_i} \lambda y_i)$ donne la stabilité par combinaison linéaire en utilisant celle de ℓ^1 et le premier point pour le troisième terme.

3/ Vous avez vu en algèbre 4 le cas des sommes finies. Il suffit de prendre la limite sur une suite exhaustive de parties finies. □

4.2 Espaces $L^p(\Omega, P)$

On prend Ω un ensemble au plus dénombrable (le cas intéressant est le cas infini). Soit $P = (p_k)_{k \in \Omega} \in \ell^1(\Omega)$ une famille sommable avec $p_k > 0$, pour tout $k \in \Omega$ (quitte à rétrécir Ω on peut se ramener à ce cas à partir du cas si on a des $p_k = 0$). On suppose aussi $\|P\|_1 = 1$.

P servira à représenter une probabilité à partir du chapitre suivant. On fixe $p \in [1, +\infty[$. (le plus souvent on aura $p \in \mathbb{N}^*$).

Définition 28. Une famille $(z_i)_{i \in \Omega}$ de nombres complexes ou réels est dite de puissance p sommable par rapport à la probabilité P si la famille $(|z_i|^p p_i)_{i \in \Omega}$ est sommable. On note $L^p(\Omega, P, \mathbb{K})$ l'ensemble des familles d'éléments de \mathbb{K} de puissance p sommable par rapport à la probabilité P .

On note

$$\|z\|_{L^p} = \left(\sum_{i \in \Omega} |z_i|^p p_i \right)^{1/p}.$$

Vous verrez en L3 que c'est une norme sur l'espace vectoriel $L^p(\Omega, P, \mathbb{K})$. Ce résultat est admis ici.

Théorème 41. $(L^p(\Omega, P, \mathbb{K}), \|\cdot\|_{L^p})$ est un e.v.n. pour tout $p \in [1, +\infty[$.

Preuve des cas $p = 1$. On voit que c'est un sous-espace vectoriel de l'ensemble des fonctions \mathbb{K}^Ω . D'abord, la famille nulle est sommable et de plus si $\lambda, \mu \in \mathbb{K}$, $(a_i p_i), (b_i p_i)$ des familles sommables, comme $\ell^1(\Omega)$ est un e.v. $(\lambda a_i p_i + \mu b_i p_i)$ l'est aussi, donc $L^1(\Omega, P, \mathbb{K})$ est stable par combinaison linéaire et est donc un sous-espace vectoriel de \mathbb{K}^Ω , puisqu'il contient aussi la famille nulle (0).

Le cas $p=2$ est similaire aux sections précédentes (laissé en exo). □

En général si il n'y a pas de confusion avec les espaces ℓ^p on note simplement $\|\cdot\|_p$ au lieu de $\|\cdot\|_{L^p}$. On note $L^\infty(\Omega, P, \mathbb{K}) = \ell^\infty(\Omega, \mathbb{K})$

(cela ne dépend pas de P sous la condition $p_k > 0$ pour tout k).

Proposition 42. *On a l'inclusion pour $1 \leq p \leq q < \infty$,*

$$L^\infty(\Omega, P, \mathbb{K}) \subset L^q(\Omega, P, \mathbb{K}) \subset L^p(\Omega, P, \mathbb{K}).$$

Attention au sens inverse de l'inclusion par rapport au cas des espace ℓ^p . (ici vue l'inclusion des suites bornées seules les grands indices de suites importent, alors que dans la section précédente, seules les petites valeurs importaient)

Démonstration. La première inclusion est simple si $z = (z_i)$ bornée (par $C > 0$) alors aussi $(|z_i|^q)$ et la famille $(|z_i|^q p_i)$ est encore sommable, avec $\|(|z_i|^q p_i)\|_1 \leq C \|P\|_1$ en passant au sup sur les inégalités pour les sommes finies.

Pour la deuxième inclusion, si $z \in L^q(\Omega, P, \mathbb{K})$ alors c'est aussi le cas de $y = (1 + |z_i|)$ (vue la propriété d'espace vectoriel et par monotonie de l'exponentielle : $(1 + |z_i|)^p \leq (1 + |z_i|)^q$ donc :

$$\sum_{i \in \Omega} |z_i|^p p_i \leq \sum_{i \in \Omega} (1 + |z_i|)^p p_i \leq \sum_{i \in \Omega} (1 + |z_i|)^q p_i < \infty$$

d'où $z \in L^p(\Omega, P, \mathbb{K})$. □

Chapitre 4

Espaces probabilisés

1 Expérience aléatoire, Introduction

On appelle expérience aléatoire toute expérience, provoquée ou non, reproductible, mais dont la répétition, dans des conditions supposées être les mêmes pour un observateur ou un expérimentateur, conduit à un ensemble de résultats dont aucun ne peut être prévu à l'avance. Cette définition se rattache à la notion de phénomènes aléatoires qui se manifestent toujours avec un certain degré d'indétermination, ce qui empêche de prévoir exactement le résultat de leurs observations, par opposition aux phénomènes déterministes régis par des lois qui en déterminent le déroulement et le résultat. On peut cependant remarquer que tout phénomène comporte une part d'aléatoire qui traduit l'impossibilité de connaître ou de maîtriser tous les paramètres qui le définissent. En fait, notre environnement comporte de nombreux phénomènes aléatoires qui interviennent dans toutes les branches de l'activité humaine. D'où la nécessité et l'importance de disposer et de développer des moyens et des techniques qui aident à comprendre ces phénomènes, à déceler et dégager des propriétés de régularité et des tendances quantifiées afin de maîtriser, prévoir et agir. A travers cette quantification, on obtient un calcul déterministe décrivant les phénomènes aléatoires.

Deux approches s'imposent et se complètent : L'observation et la modélisation. Mais un traitement qui ne s'appuie pas sur un modèle solide atteint vite ses limites et peut conduire à des résultats incohérents appelés paradoxes. La modélisation fournit un cadre d'étude fondé sur une théorie cohérente qui possède et permet de développer des outils efficaces et rationnels. L'objet fondamental de cette modélisation est l'espace probabilisé qui est un triplet (Ω, \mathcal{T}, P) dont les termes sont présentés dans ce chapitre.

Par ailleurs, l'utilisation d'aléa permet la réalisation d'algorithmes parfois plus efficaces ou l'obtention plus aisée d'algorithmes efficaces. On proposera donc la simulation de la plupart des objets introduits en utilisant un générateur de nombres aléatoires comme point de départ.

2 L'espace des réalisations

L'espace des réalisations Ω d'une expérience aléatoire est l'ensemble des résultats possibles de cette expérience. Cet espace est aussi appelé univers, référentiel, ensemble fondamental ou espace des épreuves.

Les éléments $\omega \in \Omega$ sont appelés réalisations de l'expérience en question.

L'ensemble Ω peut être fini, infini dénombrable ou infini non dénombrable.

Selon le point de vue et l'objectif de l'étude, on peut considérer des ensembles différents pour une

même expérience. Par exemple, on peut s'intéresser aux six nombres amenés par le lancer d'un dé ou uniquement à leur parité. Dans le premier cas, on prendra $\Omega = \{1, 2, 3, 4, 5, 6\}$ et dans le second $\Omega = \{0, 1\}$.

Selon les cas, l'ensemble Ω peut être plus ou moins simple ou plus ou moins compliqué. Parfois, il n'est pas aisé de définir explicitement Ω mais il est fondamental de pouvoir le faire, au moins dans les cas simples. Lorsqu'une idée correcte de la nature et de l'importance de Ω est acquise, on peut utiliser efficacement le modèle avec une définition implicite de cet espace.

3 Évènements et tribu d'évènements

Dans l'étude d'une expérience aléatoire, on s'intéresse à des ensembles de réalisations appelés évènements. Un évènement lié à une expérience aléatoire et à son espace Ω s'identifie donc à une partie de Ω .

Les évènements élémentaires s'identifient aux singletons $\{\omega\}$.

Définition 29. Soit $A \subset \Omega$ un évènement et $\omega \in \Omega$ une réalisation de l'expérience. On dit que l'évènement A s'est réalisé lorsque $\omega \in A$ et ne s'est pas réalisé lorsque $\omega \notin A$.

Remarque 16. Tout évènement étant une partie de Ω , une correspondance s'établit alors entre le langage naturel des évènements et celui des sous-ensembles. Nous en donnons ici quelques exemples.

- L'espace Ω représente l'évènement certain et l'évènement impossible correspond à l'ensemble vide \emptyset .
- L'évènement contraire à l'évènement A est représenté par son complémentaire A^c .
- Deux évènements A et B sont dits *incompatibles* lorsqu'ils sont disjoints, i.e. lorsque $A \cap B = \emptyset$.

Définition 30. On appelle *système complet d'évènements* (SCE) toute suite finie ou infinie $(A_n)_{n \geq 1}$ d'évènements deux à deux incompatibles telle que $\cup_{n \geq 1} A_n = \Omega$. (On parle aussi parfois de partition si ils sont tous non vides).

Nous verrons plus loin que cette notion est très utile car elle ramène à des situations où on dispose de plus d'informations pour traiter les évènements étudiés.

Pour un espace de réalisation fini, on pouvait prendre simplement $\mathcal{P}(\Omega)$ comme ensemble d'évènements, qui était fini, ce qui limitait aux unions finies les unions envisageables dans la définition d'une probabilité. Dans le cas infini dénombrable, on a déjà $\mathcal{P}(\Omega)$ non dénombrable et la situation est plus compliquée. Bien qu'on pourra encore prendre $\mathcal{P}(\Omega)$ comme ensemble d'évènements, on introduit le formalisme qui sera utilisé en L3. Pour l'instant, on observera juste que la définition d'un ensemble de tous les évènements pertinents pour l'expérience, qui portera le nom de tribu, est juste ce qui est nécessaire pour définir une notion de probabilité.

Définition 31. La famille \mathcal{T} des évènements constitue une partie de $\mathcal{P}(\Omega)$ ayant la structure d'une tribu qui, par définition, possède les propriétés suivantes :

1. $\emptyset \in \mathcal{T}, \Omega \in \mathcal{T}$.
2. Si $A \in \mathcal{T}$ alors $A^c \in \mathcal{T}$.
3. Pour toute suite finie ou infinie (dénombrable) $(A_n)_{n \geq 1}$ de parties de \mathcal{T} , alors $\cup_{n \geq 1} A_n \in \mathcal{T}$.

Par construction (ou hypothèse), la tribu \mathcal{T} contient tous les évènements auxquels on s'intéresse lors de l'étude d'une expérience aléatoire. La structure de tribu est importante pour l'efficacité du modèle et pour donner un sens aux différentes opérations que l'on est amené à faire sur les évènements.

En pratique, on peut supposer que toutes ces propriétés sont satisfaites sans référence explicite à la tribu, d'autant plus que, *lorsque Ω est fini ou dénombrable, on peut prendre pour tribu \mathcal{T} la famille $\mathcal{P}(\Omega)$ de toutes les parties de Ω .*

Cependant, il faut toujours avoir conscience de la tribu retenue pour l'étude. En effet, lorsque Ω est infini non dénombrable (par exemple $\Omega = \mathbb{R}$) la tribu $\mathcal{P}(\Omega)$ est trop grande et soulève des difficultés d'ordre technique. Il y a alors un choix judicieux à faire et la tribu considérée sera assez riche pour contenir tous les évènements qui interviennent lors de l'étude entreprise.

4 Probabilité sur une tribu d'évènements

Ayant défini les évènements et leur famille, on est amené à quantifier la vraisemblance de chacun d'eux et à évaluer ses chances de réalisation au cours d'une expérience aléatoire. Ceci se fait en associant à chaque évènement un nombre compris entre 0 et 1 appelé sa probabilité, en respectant quelques règles inspirées par l'intuition. Ceci nous conduit à la définition suivante

Définition 32 (Définition d'une probabilité). On appelle *probabilité* sur une tribu \mathcal{T} toute application $P : \mathcal{T} \rightarrow [0, 1]$ ayant les propriétés suivantes :

1. $P(\emptyset) = 0, P(\Omega) = 1$.
2. Pour toute suite finie ou infinie (dénombrable) $(A_n)_{n \geq 1}$ d'évènements deux à deux incompatibles c'est à dire les A_n deux à deux disjoints),

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

La propriété 2., appliquée à une suite finie, est l'additivité d'une probabilité et, appliquée à une suite infinie, est la σ -additivité d'une probabilité. Une probabilité est aussi appelée *loi* ou *distribution*.

5 Premières propriétés d'une probabilité

Malgré la simplicité apparente de sa définition, une probabilité possède des propriétés très intéressantes qu'il est essentiel de connaître car elles nous permettent d'accéder à la probabilité d'évènements variés et très complexes.

Notons que tous les évènements considérés dans les relations suivantes appartiennent à la tribu \mathcal{T} sur laquelle la probabilité P est définie.

Proposition 43. 1. $P(A^c) = 1 - P(A)$.

2. Si $A \subset B$ alors $P(B) = P(A) + P(B - A)$. En particulier, si $A \subset B$ alors $P(A) \leq P(B)$.

3. Si $A \cap B = \emptyset$, alors $P(A \cup B) = P(A) + P(B)$. Dans le cas général, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Plus généralement encore,

4.

$$P\left(\bigcup_{k=1}^n A_k\right) = \sum_{k=1}^n (-1)^{k-1} \sum_{1 \leq i_1 < \dots < i_k \leq n} P(A_{i_1} \cap \dots \cap A_{i_k}).$$

(Formule du crible ou de Poincaré).

5. Si $(A_n)_{n \geq 1}$ est une suite croissante,

$$P\left(\bigcup_{n \geq 1} A_n\right) = \lim_{n \rightarrow \infty} P(A_n) = \sup_{n \geq 1} P(A_n).$$

6. Si $(A_n)_{n \geq 1}$ est une suite décroissante,

$$P\left(\bigcap_{n \geq 1} A_n\right) = \lim_{n \rightarrow \infty} P(A_n) = \inf_{n \geq 1} P(A_n).$$

Démonstration. cf cours pour la plupart des propriétés : Montrons par récurrence sur $n \geq 2$ la formule de Poincaré.

Pour $n = 2$, c'est le 2 :

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2).$$

Supposons le résultat au rang n et montrons le rang $n + 1$: En posant $B = \bigcup_{k=1}^n A_k$, on a $\bigcup_{k=1}^{n+1} A_k = B \cup A_{n+1}$ donc

$$P\left(\bigcup_{k=1}^{n+1} A_k\right) = P(B) + P(A_{n+1}) - P(B \cap A_{n+1}).$$

Or par distributivité des intersections et des unions : $B \cap A_{n+1} = \bigcup_{k=1}^n A_k \cap A_{n+1}$ et si on pose $B_k = A_k \cap A_{n+1}$ on a $B_{i_1} \cap \dots \cap B_{i_k} = A_{i_1} \cap \dots \cap A_{i_k} \cap A_{n+1}$. De l'hypothèse de récurrence, on déduit :

$$P\left(\bigcup_{k=1}^n A_k\right) = \sum_{k=1}^n (-1)^{k-1} \sum_{1 \leq i_1 < \dots < i_k \leq n} P(A_{i_1} \cap \dots \cap A_{i_k}).$$

$$\begin{aligned} P\left(\bigcup_{k=1}^n (A_k \cap A_{n+1})\right) &= \sum_{k=1}^n (-1)^{k-1} \sum_{1 \leq i_1 < \dots < i_k \leq n} P(B_{i_1} \cap \dots \cap B_{i_k}) \\ &= \sum_{k=1}^n (-1)^{k-1} \sum_{1 \leq i_1 < \dots < i_k \leq n} P(A_{i_1} \cap \dots \cap A_{i_k} \cap A_{n+1}). \end{aligned}$$

Donc

$$\begin{aligned} P\left(\bigcup_{k=1}^{n+1} A_k\right) &= \sum_{k=1}^n (-1)^{k-1} \sum_{1 \leq i_1 < \dots < i_k \leq n} P(A_{i_1} \cap \dots \cap A_{i_k}) + P(A_{n+1}) \\ &\quad - \sum_{k=1}^n (-1)^{k-1} \sum_{1 \leq i_1 < \dots < i_k \leq n} P(A_{i_1} \cap \dots \cap A_{i_k} \cap A_{n+1}). \end{aligned}$$

Or on peut décomposer la somme voulue de la même manière en 3 sommes (celles sans A_{n+1} , celle avec un seul terme et celle avec A_{n+1} et d'autres termes :

$$\begin{aligned} \sum_{k=1}^{n+1} (-1)^{k-1} \sum_{1 \leq i_1 < \dots < i_k \leq n+1} P(A_{i_1} \cap \dots \cap A_{i_k}) &= \sum_{k=1}^n (-1)^{k-1} \sum_{1 \leq i_1 < \dots < i_k \leq n} P(A_{i_1} \cap \dots \cap A_{i_k}) \\ &\quad + P(A_{n+1}) + \sum_{k=2}^{n+1} (-1)^k \sum_{1 \leq i_1 < \dots < i_k = n+1} P(A_{i_1} \cap \dots \cap A_{i_{k-1}} \cap A_{n+1}) \end{aligned}$$

Ce qui conclut l'induction. □

Définition 33. Un évènement A est dit *presque certain* (ou *presque sûr*) si $P(A) = 1$ et *presque impossible* si $P(A) = 0$.

Définition 34. On appelle système presque complet d'évènements (SPCE) toute suite finie ou infinie (dénombrable) $(A_n)_{n \geq 1}$ d'évènements deux à deux incompatibles telle que

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n) = 1.$$

Proposition 44. Soit $(A_n)_{n \geq 1}$ un SPCE. Alors, pour tout évènement B , on a

$$P(B) = \sum_{n \geq 1} P(A_n \cap B).$$

Il est important de noter qu'il n'existe pas de probabilité unique et prédéfinie. Le choix et la construction d'une probabilité pour modéliser une expérience aléatoire doit tenir compte de ses propriétés spécifiques et refléter autant que possible la réalité du phénomène. Un mauvais choix conduit à un modèle qui a sa cohérence interne mais complètement inadapté au phénomène qu'il est censé décrire. Il fournit des résultats en contradiction avec les observations que l'on peut faire. Il est donc inutile.

Le paragraphe suivant donne la méthode générale de construction d'une probabilité sur un ensemble fini ou dénombrable.

6 Probabilité sur un ensemble fini ou dénombrable

Soit Ω un ensemble fini ou dénombrable. La donnée d'une probabilité P sur $\mathcal{P}(\Omega)$ équivaut à la donnée d'une fonction $p : \Omega \rightarrow [0, 1]$ telle que $\sum_{\omega \in \Omega} p(\omega) = 1$. La correspondance se fait de la manière suivante : Lorsque P est donnée, on définit p par les relations $p(\omega) = P(\{\omega\})$ et lorsque p est donnée, on définit P par la relation

$$P(A) = \sum_{\omega \in A} p(\omega).$$

Cette construction est explicitée davantage dans ce qui suit.

6.1 Cas des ensembles finis

Une probabilité sur (la famille de toutes les parties d') un ensemble fini $\Omega = \{\omega_1, \dots, \omega_n\}$ est complètement déterminée par la donnée de n nombres positifs $(p_k)_{1 \leq k \leq n}$ tels que $\sum_{k=1}^n p_k = 1$. Le nombre p_k représente alors la probabilité de la réalisation ω_k et, pour tout $A \subset \Omega$, on a $P(A) = \sum_{k \in A} p_k$.

6.2 Cas d'un ensemble dénombrable

Une probabilité sur (la famille de toutes les parties d') un ensemble infini dénombrable $\Omega = \{\omega_n\}_{n \geq 1}$ est complètement déterminée par la donnée d'une suite $(p_n)_{n \geq 1}$ de nombres positifs telle que $\sum_{n=1}^{\infty} p_n = 1$. Le nombre p_n représente alors la probabilité de la réalisation ω_n et, pour tout $A \subset \Omega$, on a

$$P(A) = \sum_{n \in A} p_n.$$

Il n'existe pas de probabilité uniforme sur un ensemble infini dénombrable.

6.3 Quelques lois usuelles sur un ensemble fini ou dénombrable

La loi uniforme $\mathcal{U}(n)$, de paramètre $n \in \mathbb{N}^*$ a été présentée ci-dessus. On a vu au chapitre 1 la loi binomiale $B(n, p)$ et la loi hypergéométrique $H(N, n, p)$.

Définition 35. La loi de Poisson $P(\lambda)$, de paramètre $\lambda > 0$. C'est la loi sur \mathbb{N} définie par les nombres $p_n = e^{-\lambda} \frac{\lambda^n}{n!}$. C'est la loi du nombre de succès associé à un phénomène discret qui se déroule dans le temps ou dans l'espace. C'est la limite de $B(n, \frac{\lambda}{n})$ quand $n \rightarrow \infty$ (cf TD).

Définition 36. La loi géométrique $G(p)$, de paramètre $p \in]0, 1[$. C'est la loi sur \mathbb{N}^* définie par les nombres $p_n = pq^{n-1}$ où $q = 1 - p$.

C'est la loi d'attente du premier succès dans une succession infinie dénombrable d'essais identiques et indépendants, lorsque la probabilité du succès à chaque essai est p . (cf chapitre suivant).

Définition 37. La loi géométrique $G_0(p)$, de paramètre $p \in]0, 1[$. C'est la loi sur \mathbb{N} définie par les nombres $p_n = pq^n$ où $q = 1 - p$.

C'est la loi du nombre d'échecs précédant le premier succès dans une succession infinie dénombrable d'essais identiques et indépendants, lorsque la probabilité du succès à chaque essai est p .

Chapitre 5

Probabilité conditionnelle. Indépendance

1 Probabilité conditionnelle

Dans la modélisation d'une expérience aléatoire, l'espace probabilisé tient compte des informations qui apparaissent ou qui modifient le déroulement de cette expérience. Ainsi, l'information concernant la réalisation d'un évènement A sera présente dans la définition d'une nouvelle probabilité appropriée. Ceci conduit à la notion de probabilité conditionnelle qui permet par ailleurs de développer des outils très utiles pour le traitement des expériences aléatoires.

Définition 38. Soit (Ω, \mathcal{T}, P) un espace probabilisé et A un évènement tel que $P(A) > 0$. La probabilité conditionnelle sachant A , notée $P(\cdot|A)$, est la probabilité sur \mathcal{T} définie par la relation

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

On parle de "probabilité de B sachant A ".

On vérifie qu'il s'agit bien d'une probabilité qui restreint l'univers à A pour tenir compte de sa réalisation. En effet $P(\emptyset|A) = 0/P(A) = 0$, $P(\Omega|A) = P(A)/P(A) = 1$ et si B_n sont disjoints, de même $B_n \cap A$ sont disjoints donc

$$P\left(\bigcup_{n=1}^{\infty} B_n|A\right) = \frac{1}{P(A)}P\left(\bigcup_{n=1}^{\infty} (B_n \cap A)\right) = \frac{1}{P(A)} \sum_{n=1}^{\infty} P(B_n \cap A) = \sum_{n=1}^{\infty} P(B_n|A).$$

Les propriétés qui suivent sont en fait des présentations différentes de la définition, mais elles sont très utiles.

Proposition 45 (Formule des probabilités composées). *Soit A et B deux évènements de probabilité non nulle. Alors, on a*

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

Plus généralement, si $(A_k)_{1 \leq k \leq n}$, $n \geq 2$ est une suite finie d'évènements telle que $P(\bigcap_{1 \leq k \leq n-1} A_k) > 0$, alors

$$P\left(\bigcap_{k=1}^n A_k\right) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n | \bigcap_{1 \leq k \leq n-1} A_k).$$

Démonstration. La première formule vient directement de la définition de la probabilité conditionnelle comme fraction. La deuxième formule s'obtient par récurrence sur n . Si $n = 2$, c'est la formule précédente. Si on suppose l'étape $n - 1$, au rang n on obtient en appliquant la formule précédente ($n = 2$) puis la formule de récurrence :

$$P\left(\bigcap_{k=1}^n A_k\right) = P\left(\bigcap_{k=1}^{n-1} A_k \cap A_n\right) = P\left(\bigcap_{k=1}^{n-1} A_k\right)P(A_n | \bigcap_{k=1}^{n-1} A_k) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n | \bigcap_{1 \leq k \leq n-1} A_k)$$

□

Proposition 46 (Théorème des probabilités totales). *Soit $(A_n)_{n \geq 1}$ un système presque complet d'évènements, tous de probabilité non nulle. Alors, pour tout évènement B , on a*

$$P(B) = \sum_{n=1}^{\infty} P(B|A_n)P(A_n).$$

On a bien sûr le même résultat dans le cas des suites finies.

Démonstration. Comme les $P(A_n) > 0$, la formule est équivalente (par définition des probabilités conditionnelles) à :

$$P(B) = \sum_{n=1}^{\infty} P(B \cap A_n).$$

Or, on sait que si $A = \bigcup_{n \geq 1} A_n$, on a par définition d'un système presque complet, $A_n, \Omega - A$ forment une partition de Ω et $P(\Omega - A) = 0$. En particulier en prenant l'intersection avec B : $B = (B - A) \cup \bigcup_{n \geq 1} (B \cap A_n)$ est une union disjointe d'où par l'axiome de σ -additivité :

$$P(B) = P(B - A) + \sum_{n=1}^{\infty} P(B \cap A_n).$$

Or $P(B - A) \leq P(\Omega - A) = 0$ d'où le résultat. □

Exemple 24. On note (N_1, N_2) , le résultat du lancer d'une paire de dés équilibrés à n faces. $\Omega = \{1, \dots, n\}^2$ avec la probabilité uniforme. On cherche la probabilité de l'évènement $N_1 > N_2$. Par la formule :

$$P(N_1 > N_2) = \sum_{k=1}^n P(N_1 > N_2 | N_1 = k)P(N_1 = k).$$

Par l'uniformité $P(N_2 = k) = \frac{1}{n}$ et

$$P(N_1 > N_2 | N_1 = k) = P(k > N_2 \cap N_1 = k) / P(N_1 = k) = P(k > N_2) = \frac{k-1}{n}$$

(ce qui sera justifié en vérifiant que les lancers de N_1 et N_2 sont indépendants.) Donc :

$$P(N_1 > N_2) = \sum_{k=1}^n \frac{k-1}{n^2} = \frac{n(n-1)}{2n^2} = \frac{n-1}{2n}.$$

Pour $n = 6$, on obtient $5/12$.

Proposition 47 (Formule de Bayes). Soit $(A_n)_{n \geq 1}$ un système presque complet d'évènements, tous de probabilité non nulle et B un évènement de probabilité non nulle également. Alors, pour tout $k \geq 1$,

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{n \geq 1} P(B|A_n)P(A_n)}.$$

Démonstration. C'est juste la définition

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

On combine avec la formule des probabilités totales pour le dénominateur. □

D'après l'interprétation qu'on peut lui donner, cette formule est aussi connue sous le nom de formule des probabilités des causes.

Exemple 25. Un laboratoire a mis au point un test pour déceler des souris malades. Des essais prouvent que :

- 96 fois sur 100, le test donne un résultat positif quand la souris est effectivement malade.
- 94 fois sur 100, le test donne un résultat négatif quand la souris n'est pas malade.

Dans une population de souris comprenant 3 % de malades, on pratique le test sur une souris choisie au hasard et on constate que le test donne un résultat positif. Quelle est la probabilité que la souris soit malade ?

On pose les évènements T , le test est positif, et M , la souris est malade. L'énoncé nous apprend que $P(T|M) = 0,96$ et $P(T^c|M^c) = 0,94$, donc $P(T|M^c) = 0,06$. Par ailleurs on a $P(M) = 3/100$, la formule de Bayes nous permet de calculer la probabilité qu'une souris soit malade sachant que le test a été positif :

$$P(M|T) = \frac{P(T|M)P(M)}{P(T|M)P(M) + P(T|M^c)P(M^c)} = \frac{0,96 \times 0,03}{0,96 \times 0,03 + 0,06 \times 0,97} = \frac{288}{870} \simeq 0,331.$$

Bien que le test semblait efficace, il garantit très peu le caractère malade si il est positif. Cela vient de ce que $P(M) \ll P(M^c) \sim 1$. Si on n'a pas pour compenser $P(T|M^c) \ll P(M)$ (même avec $P(T|M) \sim 1$) mais comme dans l'énoncé $P(T|M^c) \simeq 2P(M)$ on obtient donc $P(M|T) \simeq \frac{P(M)}{P(M)+P(T|M^c)} \simeq \frac{1}{3}$.

2 Indépendance

Intuitivement, on peut dire que deux évènements A et B sont indépendants si la probabilité de l'un d'eux ne change pas quand on sait que l'autre s'est réalisé. En traduisant cette idée avec la définition de la probabilité conditionnelle, on obtient la définition suivante :

Définition 39. Soit (Ω, \mathcal{T}, P) un espace probabilisé. Deux évènements A et B sont indépendants (pour la probabilité P) lorsque

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

Autrement dit si $P(B) > 0$: $P(A|B) = P(A)$. On a la propriété facile permettant de rassembler l'indépendance par union qu'on laisse en exo :

Lemme 48. Soit $(A_k)_{k \geq 0}$ des évènements formant une partition au plus dénombrable de A et $(B_l)_{l \geq 0}$ des évènements formant une partition au plus dénombrable de B . Si pour tout k, l A_k , et B_l sont indépendants alors A et B sont indépendants.

Exemple 26. On note (N_1, N_2) , le résultat du lancer d'une paire de dés équilibrés à n faces. $\Omega = \{1, \dots, n\}^2$ avec la probabilité uniforme. Montrons que les évènements $A = \{N_1 = i\}$ et $B = \{N_2 = j\}$, $i, j \in \{1, \dots, n\}$ sont indépendants (On dira que les v.a. N_1, N_2 sont indépendantes). En effet on obtient

$$P(A \cap B) = P((N_1, N_2) = (i, j)) = \frac{1}{n^2}.$$

$$P(A) = \sum_{j=1}^n P((N_1, N_2) = (i, j)) = \frac{n}{n^2} = \frac{1}{n}$$

et de même $P(B) = \frac{1}{n}$ d'où :

$$P(A \cap B) = \frac{1}{n^2} = P(A)P(B).$$

Définition 40 (Indépendance). . Deux v.a. X et Y sur un espace fini Ω sont dites indépendantes si l'on a, pour tous $x \in X(\Omega), y \in Y(\Omega)$, $\{X = x\}$ et $\{Y = y\}$ sont indépendants, c'est à dire :

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

Cette définition se généralise de la manière suivante :

Définition 41. Soit $n \in \mathbb{N}^*$. Les évènements de la suite $(A_k)_{1 \leq k \leq n}$ sont *deux à deux indépendants* si pour tout $(i, j) \in \llbracket 1, n \rrbracket^2$ tel que $i \neq j$,

$$P(A_i \cap A_j) = P(A_i)P(A_j).$$

La notion la plus utile est la notion suivante :

Définition 42. Soit $n \in \mathbb{N}^*$. Les évènements de la suite $(A_k)_{1 \leq k \leq n}$ sont *indépendants* (ou mutuellement indépendants) si pour toute partie $J \subset \llbracket 1, n \rrbracket$,

$$P(\cap_{k \in J} A_k) = \prod_{k \in J} P(A_k).$$

De même, sur un espace fini Ω , p des variables aléatoires $(X_k)_{1 \leq k \leq n}$ sont *indépendantes* si pour tous $x_k \in X_k(\Omega)$ les familles d'évènements $(\{X_k = x_k\})_{1 \leq k \leq n}$ sont *indépendants*.

Définition 43. Les évènements de la suite infinie $(A_k)_{k \geq 1}$ sont *indépendants* si, pour tout $n \in \mathbb{N}^*$, les évènements des suites $(A_k)_{1 \leq k \leq n}$ le sont.

Remarque 17. 1. La notion d'indépendance n'est pas absolue et dépend essentiellement de la probabilité définie sur la tribu des évènements. Deux évènements définis par les mêmes termes dans deux modèles différents peuvent être indépendants dans l'un des modèles sans l'être dans l'autre.

2. Un évènement presque impossible ou presque certain est indépendant de tout évènement et un évènement est indépendant de lui-même si et seulement s'il est presque certain ou presque impossible. Cette propriété montre à nouveau que la notion d'indépendance dépend de la probabilité considérée.

3. Il ne faut pas confondre les notions d'indépendance et d'incompatibilité, la dernière étant déjà une notion spécifique des évènements et elle contredit en quelque sorte la première. En fait, deux évènements incompatibles sont indépendants si et seulement si l'un d'eux est presque impossible.
4. L'indépendance des évènements de l'un des quatre couples $(A; B); (A^c; B^c); (A; B^c)$ ou $(A^c; B)$ est équivalente à l'indépendance des évènements de chacun de ces couples. Ceci se généralise au cas de n évènements.
5. L'indépendance deux à deux de n évènements se traduit par $\binom{n}{2} = \frac{n(n-1)}{2}$ égalités qui sont déjà présentes dans la définition de l'indépendance de n évènements. Mais celle-ci requiert $2^n - n - 1$ égalités. Ainsi, l'indépendance entraîne l'indépendance deux à deux mais la réciproque est fausse.

Exemple 27. On lance 25 dés équilibrés indépendamment. Calculons la probabilité de l'évènement A "avoir au moins un 1". L'évènement contraire est A^c n'avoir aucun 1 qui est l'intersection des évènements indépendants, ne pas avoir un 1 au i -ème lancer $N_i \neq 1$, donc :

$$P(A) = 1 - \prod_{i=1}^{25} P(N_i \neq 1) = 1 - \left(\frac{5}{6}\right)^{25}.$$

2.1 Retour sur l'interprétation des lois géométriques et binomiales

Exemple 28 (Loi Géométrique). On considère une suite (X_i) de lancers de pile ou face biaisés indépendants avec $P(X_i = P) = p \in [0, 1]$. On se demande qu'elle est la distribution de probabilité de l'indice N du premier pile obtenu. (On dira très vite que N est une variable aléatoire)

L'évènement $\{N = n\} = \{X_1 = F\} \cap \dots \cap \{X_{n-1} = F\} \cap \{X_n = P\}$, et comme ces évènements sont indépendants par hypothèse, on obtient :

$$P(N = n) = P(X_1 = F) \dots P(X_{n-1} = F) P(X_n = P) = (1 - p)^{n-1} p.$$

On retrouve les probabilités de la loi géométrique $G(p)$.

Exemple 29 (Application à la simulation des probabilités conditionnelles). Dans la suite, on supposera souvent que

Si l'on a une procédure donnant un résultat aléatoire, la répétition de la procédure donne des tirages indépendants.

Montrons le résultat suivant.

Lemme 49. Si on suit l'algorithme

Répéter Epreuve

$X :=$ Résultat de l'épreuve

Jusqu'à (Condition réalisée) On obtient $P(X \in B) = P(\text{"Résultat de l'épreuve dans } B" | \text{Condition})$.

Démonstration. Soit B_n l'évènement "le résultat est dans B à l'épreuve n ". Soit C_n l'évènement "la condition C se réalise à l'épreuve n ".

Alors

$$\{X \in B\} = \bigcup_{n=1}^{\infty} C_1^c \cap \dots \cap C_{n-1}^c \cap C_n \cap B_n,$$

donc comme les évènements sont disjoints et par indépendances des lancers :

$$P(X \in B) = \sum_{n=1}^{\infty} P(C_1^c) \dots P(C_{n-1}^c) P(C_n \cap B_n) = \frac{P(C \cap B)}{1 - P(C^c)} = P(B|C).$$

□

Par exemple,

Si on effectue :

Répéter

$U := \text{dé}(8)$

Jusqu'à ($U \leq 6$)

$X := U$

On obtient la simulation d'un dé à six faces à partir d'un dé à 8 faces.

En effet pour $k = 1, \dots, 6$,

$$P(X = k) = P(U = k | U \leq 6) = \frac{1/8}{6/8} = \frac{1}{6}.$$

Exemple 30. (Loi binomiale) On considère comme à l'exemple 28 une suite (X_i) de lancers de pile ou face biaisés indépendants avec $P(X_i = P) = p \in [0, 1]$.

On note N_n le nombre de piles obtenus parmi les n premiers lancers.

$$\{N_n = k\} = \cup_{\mathcal{P} \in P_k(\llbracket 1, n \rrbracket)} \left(\left(\cap_{i \in \mathcal{P}} \{X_i = P\} \right) \cap \left(\cap_{i \notin \mathcal{P}} \{X_i = F\} \right) \right).$$

Par union disjointe et indépendance, on déduit

$$P(N_n = k) = \sum_{\mathcal{P} \in P_k(\llbracket 1, n \rrbracket)} P(X_i = P)^k P(X_i = F)^{n-k} = \text{Card}(P_k(\llbracket 1, n \rrbracket)) p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}.$$

On retrouve la loi binomiale.

Chapitre 6

Variables aléatoires discrètes

1 Introduction

Le plus souvent, les questions liées à l'étude d'une expérience aléatoire s'expriment en termes d'une fonction numérique ou vectorielle des réalisations. Il est donc naturel de s'intéresser à ce type de fonctions et c'est ce qui nous conduit à la notion de variable et vecteur aléatoire. Nous considérons ici uniquement les variables aléatoires. A l'aide de ces variables, on dispose à la fois d'une interprétation concrète de l'objet d'étude et d'un moyen de mettre à profit des outils analytiques puissants. Leur utilité est telle qu'il est toujours recommandé, au moyen d'un codage approprié s'il en est besoin, d'associer une variable ou un vecteur aléatoire aux questions considérées. L'étude d'une variable ou vecteur aléatoire porte principalement sur sa loi (sa répartition) et ses paramètres caractéristiques.

2 Exemples

On rappelle les exemples déjà considéré 19 et 20.

Exemple 31. Le nombre de portables vendus dans une journée dans un magasin est une variable aléatoire.

Exemple 32. Un schéma général largement utilisé est la répétition d'une expérience aléatoire élémentaire à deux issues appelées succès et échec. Au k -ème essai, on associe la variable aléatoire X_k qui vaut 1 lorsque l'issue correspondante est succès et qui vaut 0 dans le cas contraire. Alors, pour tout entier $n \in \mathbb{N}^*$, $S_n = \sum_{k=1}^n X_k$ est la variable aléatoire égale au nombre de succès en n essais.

Exemple 33. En choisissant au hasard un point dans le disque de centre O et de rayon R , on peut s'intéresser à la variable aléatoire D égale à la distance du point choisi au centre O du disque.

Exemple 34. La hauteur des précipitations en un mois dans une région donnée est une variable aléatoire.

Exemple 35. Le temps de service d'un client à un guichet est une variable aléatoire.

La notion de variable aléatoire et l'étude de sa répartition se font à partir de la construction d'un espace probabilisé dont le premier terme est \mathbb{R} . Ceci est précisé dans les définitions suivantes :

3 Définitions dans le cas général. Loi d'une variable aléatoire (facultatif)

Soit (Ω, \mathcal{T}, P) un espace probabilisé. On va très vite se restreindre au cas de variables aléatoires discrètes, mais on introduit quelques terminologies qui seront étudiées plus en détail en L3.

Pour définir la tribu des évènements de \mathbb{R} , on peut penser à la famille I des intervalles de \mathbb{R} mais on s'aperçoit vite que celle-ci n'est pas une tribu. Le bon choix se fait en considérant la plus petite tribu contenant les intervalles.

Définition 44. Par définition, la plus petite tribu contenant les intervalles, notée $\mathcal{B}(\mathbb{R})$ ou \mathcal{B} , est la tribu borélienne de \mathbb{R} . Un ensemble $B \subset \mathbb{R}$ tel que $B \in \mathcal{B}$ est appelé borélien de \mathbb{R} .

Dans la pratique, toute partie B de \mathbb{R} qui apparaît "de manière naturelle" dans l'étude d'une expérience aléatoire sera une partie borélienne et nous la considérons toujours comme telle sans se poser de questions sur sa nature (vous étudierez plus cette notion durant le cours de théorie de la mesure en L3). Le plus souvent, B sera un intervalle ou une réunion d'intervalles. Par défaut, \mathbb{R} sera toujours muni de sa tribu borélienne \mathcal{B} .

Définition 45. Soit $S \subset \mathbb{R}$ un borélien. La tribu borélienne de S , notée $\mathcal{B}(S)$ est constituée de tous les boréliens B de \mathbb{R} tels que $B \subset S$.

Là aussi, toute partie de S qui apparaît dans l'étude sera borélienne.

Définition 46. Une variable aléatoire (v.a.) définie sur Ω est une application $X : \Omega \rightarrow \mathbb{R}$ (ou $X : \Omega \rightarrow \overline{\mathbb{R}}$) telle que, pour tout borélien B de \mathbb{R} , l'ensemble

$$\{X \in B\} \equiv X^{-1}(B) \equiv \{\omega \in \Omega : X(\omega) \in B\}$$

est un évènement, i.e. $\{X \in B\} \in \mathcal{T}$. Pour cela, on démontre (cf L3) qu'il faut et il suffit que cette propriété soit satisfaite lorsque B est un intervalle de \mathbb{R} .

En L3 vous direz que ce type d'applications sont "mesurables" pour les tribus \mathcal{T} et \mathcal{B} .

Cette définition s'explique naturellement de la manière suivante. Du point de vue probabiliste, la caractéristique essentielle d'une v.a. est sa répartition qui permet de déterminer la probabilité avec laquelle cette variable appartient à telle ou telle région et ceci suppose que l'on soit en mesure de calculer la probabilité des ensembles signalés ci-dessus tels que $\{X \in B\}$ et pour cela il faut que ces ensembles appartiennent à la tribu des évènements sur laquelle la probabilité est définie.

Définition 47. Une application $X : \Omega \rightarrow \mathbb{C}$ est une v.a. si sa partie réelle et sa partie imaginaire le sont.

Là encore, toute application $X : \Omega \rightarrow \mathbb{R}$ ou $X : \Omega \rightarrow \mathbb{C}$ qui apparaît dans l'étude d'une expérience aléatoire sera considérée d'office comme une variable aléatoire.

Définition 48. Soit $X : \Omega \rightarrow \mathbb{R}$ une v.a. On appelle loi (distribution, ou répartition) de X la probabilité sur \mathcal{B} , notée P_X ou μ_X , définie par

$$P_X(B) = P(\{X \in B\}), B \in \mathcal{B}$$

On vérifie que P_X ainsi définie est effectivement une probabilité (car l'image inverse respecte les opérations ensemblistes comme le complémentaire et les unions dénombrables).

Soit $S = X(\Omega)$ l'ensemble des valeurs de X . Lorsque S est distinct de \mathbb{R} , on a, par additivité, pour tout $B \in \mathcal{B}$, $P_X(B) = P_X(B \cap S) + P_X(B \cap S^c) = P_X(B \cap S)$ car $P_X(B \cap S^c) = P(X \in B \cap S^c) = P(\emptyset) = 0$. Cette constatation montre que la loi de X est complètement définie par ses valeurs sur les boréliens de S .

Ainsi, on peut considérer la loi de X comme une probabilité sur S . Mais lorsque l'on fait l'étude de plusieurs v.a. à la fois, il est plus commode de considérer que toutes les lois sont définies sur \mathbb{R} .

Avec ce qui précède, on voit qu'une variable aléatoire X permet de construire un nouvel espace probabilisé $(\mathbb{R}, \mathcal{B}, P_X)$ qui va permettre d'établir toutes les caractéristiques de X . Les probabilités sont surtout l'étude des questions sur les variables aléatoires qui se ramènent à l'étude de la loi, indépendamment de la construction concrète de Ω .

Voici deux notions qui peuvent intervenir quand on est en présence de deux variables aléatoires.

Définition 49 (Egalité en loi). . Deux v.a. X et Y ayant même loi $P_X = P_Y$ sont dites équidistribuées ou égales en loi ou *identiquement distribuées*, ce que l'on peut noter $X \stackrel{d}{=} Y$. Insistons sur le fait que l'égalité en loi est distincte de l'égalité ponctuelle.

Définition 50 (Indépendance). . Deux v.a. X et Y sont dites indépendantes si l'on a, pour tous boréliens A et B de \mathbb{R} ,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

Pour cela (cf cours de L3), il faut et il suffit que cette condition soit satisfaite lorsque A et B sont des intervalles quelconques de \mathbb{R} .

Remarque 18. 1. X et Y sont donc indépendants si $\{X \in A\}$ et $\{Y \in B\}$ sont des évènements indépendants pour TOUS boréliens (ou intervalles) A et B .

2. Lorsque X et Y sont deux v.a. indépendantes, f et g deux fonctions boréliennes (i.e. des v.a. sur $(\mathbb{R}, \mathcal{B})$), alors les v.a. $f(X)$ et $g(Y)$ sont encore indépendantes.

3. Les définitions précédentes se généralisent au cas de plusieurs v.a. Une suite $(X_n)_{n \geq 1}$ de v.a. sont dites indépendantes si pour toute suite de boréliens $(A_n)_{n \geq 1}$, la suite d'évènements $(\{X_n \in A_n\})_{n \geq 1}$ est indépendante.

Les v.a. usuelles sont de deux types et pour chacun d'eux, on dispose d'outils adaptés pour faire le calcul des lois et de leurs caractéristiques. On présente d'abord le premier type, les variables discrètes, puis le second, les variables continues.

4 Variables aléatoires discrètes

Définition 51. Une variable aléatoire $X : \Omega \rightarrow \mathbb{R}$ (ou $X : \Omega \rightarrow \overline{\mathbb{R}} = \mathbb{R} \cup \{\infty, -\infty\}$ ou $X : \Omega \rightarrow \mathbb{C}$) est dite *discrète (v.a.d.)* si l'ensemble $S = X(\Omega)$ de ses valeurs est une partie finie ou dénombrable de \mathbb{R} (de $\overline{\mathbb{R}}$ ou de \mathbb{C}).

Remarque 19. Dans ce cours, le plus souvent, Ω sera au plus dénombrable et toute fonction $X : \Omega \rightarrow \mathbb{R}$ sera une v.a. discrète. Si (Ω, \mathcal{T}, P) est infini non-dénombrable, une fonction $X : \Omega \rightarrow S \subset \mathbb{R}$, à valeur dans un sous-ensemble dénombrable de \mathbb{R} , est une v.a. discrète si et seulement si pour tout $x \in X(\Omega)$, $X^{-1}(\{s\}) \equiv \{X = s\} \in \mathcal{T}$. On peut prendre cela pour définition dans ce cours. (cela vient du fait que $P(S) \subset \mathcal{B}(\mathbb{R})$ et que toutes ces parties s'écrivent comme union dénombrables de singletons)

Dans cette partie, on se limitera aux v.a.d. et, sauf mention expresse du contraire, elles seront à valeurs dans \mathbb{R} .

4.1 Loi d'une v.a.d.

La loi d'une v.a.d. X est donc la probabilité P_X , définie sur l'ensemble fini ou dénombrable $S = X(\Omega)$ de ses valeurs. D'après le résultat du chapitre 3, on voit que P_X est complètement définie par la suite des nombres positifs $p_s = P(X = s)$, lorsque s parcourt S . Bien entendu, $\sum_{s \in S} p_s = 1$. Alors, par additivité ou σ -additivité, on a pour toute partie (borélienne) $A \subset R$,

$$P(X \in A) = \sum_{s \in S \cap A} p_s.$$

et ceci définit entièrement la loi de X .

Ainsi, lorsque la v.a. est discrète, on considère que sa loi est correctement définie quand on précise l'ensemble de ses valeurs et la probabilité avec laquelle chacune de ses valeurs est prise.

Exemple 36. Dans l'exemple 20, on considère des tirages au hasard et sans remise et note X le rang d'apparition de la première boule rouge. On a $S = X(\Omega) = \llbracket 1, 5 \rrbracket$. $\text{Card}(\Omega) = \binom{10}{6}$ $\Omega = P_4(\llbracket 1, 10 \rrbracket)$ (les positions des boules rouges dans un tirage complet. La loi de X est définie par $p_1 = \frac{\binom{9}{5}}{\binom{10}{6}} = \frac{3}{5}$,

$$p_2 = \frac{\binom{8}{5}}{\binom{10}{6}} = \frac{4}{15}, p_3 = \frac{\binom{7}{5}}{\binom{10}{6}} = \frac{1}{10}, p_4 = \frac{\binom{6}{5}}{\binom{10}{6}} = \frac{1}{35} \text{ et } p_5 = \frac{\binom{5}{5}}{\binom{10}{6}} = \frac{1}{210}.$$

Exemple 37. Dans l'exemple 32, on note p la probabilité du succès à chaque essai. Alors, $S_n \sim B(n, p)$.

Dans les énoncés qui suivent, X désigne une v.a.d. dont la loi est définie par la suite des nombres $(p_s)_{s \in S}$, $p_s = P(X = s)$.

4.2 Suites de v.a. indépendantes

Définition 52 (Indépendance de n variables). Une suite finie $(X_k)_{1 \leq k \leq n}$ de v.a. discrètes est dite indépendante si l'on a, pour toutes valeurs $x_k \in X_k(\Omega)$ les événements $\{X_k = x_k\}_{1 \leq k \leq n}$ sont indépendants

Définition 53 (Indépendance d'une suite de variables). Une suite $(X_k)_{k \geq 1}$ de v.a. discrètes est dite indépendante si l'on a pour tout n les suites finies $(X_k)_{1 \leq k \leq n}$ sont indépendantes.

Il n'est pas clair qu'il existe une suite de v.a. indépendantes. On peut facilement construire des suites finies de v.a. de n Bernoulli indépendantes sur $\Omega = \{0, 1\}^n$. Pour construire une suite infinie, il faut un Ω non-dénombrable. Ceci revient à la construction de la mesure de Lebesgue sur \mathbb{R} et sera donc vu en L3. Nous admettrons par la suite cette existence.

Exemple 38. Sur $\Omega = \mathcal{P}(\llbracket 1, n \rrbracket) \simeq \{0, 1\}^n$ on note ω_A l'élément associé à la partie A , on définit la probabilité P par $P(\omega_A) = p^{\text{Card}(A)}(1-p)^{\text{Card}(A^c)}$. Comme la $\sum_{A \in \Omega} P(\omega_A) = \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = (1-p+p)^n = 1$, c'est bien une proba. On pose $X_i(\omega_A) = 1_A(i)$. Voyons que les X_i sont des variables de Bernoulli indépendantes. Par définition,

$$P(X_1 = i_1, \dots, X_n = i_n) = P(\omega_{A_i}) = p^{\sum i_j} (1-p)^{n - \sum i_j} = P(X_1 = i_1) \dots P(X_n = i_n)$$

avec A_i tel que $1_{A_i} = i$. En effet, on a bien $P(X_i = 0) = \sum_{A \in \Omega: i \notin A} P(\omega_A) = \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-1-k} = (1-p)^{n-1}$. Cette égalité donne la loi pour des variables de Bernoulli vu qu'elles sont par construction à valeur $\{0, 1\}$ L'égalité du produit pour $P(X_1 = i_1, \dots, X_n = i_n) = P(X_1 = i_1) \dots P(X_n = i_n)$ donne la propriété d'indépendance vu que c'est pour tout i_1, \dots, i_n .

4.3 Fonction de répartition (f.r.) d'une v.a.d.

Définition 54. La fonction de répartition de $X : \Omega \rightarrow \mathbb{R}$ est la fonction $F_X : \mathbb{R} \rightarrow [0, 1]$ définie par

$$F_X(x) = P(X \leq x) = \sum_{s \leq x} p_s.$$

Quand il n'y a pas de confusion à craindre, on note plus simplement F la f.r. de X .

Proposition 50. Toute fonction de répartition F_X d'une v.a.d. est croissante, continue à droite et vérifie

$$\lim_{x \rightarrow -\infty} F_X(x) = 0$$

et

$$\lim_{x \rightarrow +\infty} F_X(x) = 1.$$

De plus, pour tout $x \in \mathbb{R}$, $P(X = x) = F_X(x) - F_X(x-)$ (où $F_X(x-) = \lim_{\epsilon \rightarrow 0^+} F(x - \epsilon)$ est la limite à gauche qui existe toujours).

Démonstration. Tout est une conséquence des propriétés générales des probabilités écrites à la proposition 43.

Comme pour $x \leq y$, $\{X \leq x\} \subset \{X \leq y\}$, on obtient par croissance de P , l'inégalité $F_X(x) = P(X \leq x) \leq P(X \leq y) = F_X(y)$. Comme pour une suite décroissante $x_n \rightarrow x$, $\cap_n \{X \leq x_n\} = \{X \leq x\}$, on obtient la limite $\lim_{x_n \rightarrow x} P(X \leq x_n) = P(\cap_n \{X \leq x_n\}) = F_X(x)$. Ensuite par contraposée, si F_X n'était pas continue à droite on trouverait une suite x_n telle que cela n'est pas vérifié, d'où la continuité à droite énoncée. De même en prenant une suite croissante $x_n \rightarrow x$, on a $\cup_n \{X \leq x_n\} = \{X < x\}$ qui donne $\lim_{x_n \rightarrow x} P(X \leq x_n) = P(\cup_n \{X \leq x_n\}) = P(X < x)$. On voit par le même argument par contraposée qu'on ne peut avoir $\lim_{\epsilon \rightarrow 0^+} F(x - \epsilon) < P(X < x)$ en extrayant une suite d'où (par la borne connue par monotonie $\lim_{\epsilon \rightarrow 0^+} F(x - \epsilon) \leq P(X < x)$) on obtient :

$$F_X(x-) = \lim_{\epsilon \rightarrow 0^+} F(x - \epsilon) = P(X < x).$$

En prenant les complémentaires $P(X = x) = P(X \leq x) - P(X < x)$, on obtient la dernière égalité. En prenant $x = \pm\infty$, cela donne en cas particulier les autres résultats voulus vu $P(X < \infty) = 1, P(X < -\infty) = 0$. \square

La dernière propriété montre que la f.r. d'une v.a.d. détermine complètement la loi de cette variable.

Divers paramètres et fonctions permettent de caractériser la loi d'une v.a.d. On commence par la moyenne qui donne une idée de l'ordre de grandeur de la variable en question.

4.4 Moyenne ou Espérance ou Moment d'ordre 1

Définition 55. La v.a.d. X est dite d'ordre 1 lorsque la série $\sum_{s \in S} |s|P(X = s)$ est convergente. Dans ce cas, on appelle Moyenne ou Espérance ou Moment d'ordre 1 de X la quantité

$$m = E[X] = \sum_{s \in S} sP(X = s).$$

Remarque 20. Avec les notations du chapitre 2, si on regarde P_X comme probabilité sur $S = X(\Omega)$. X est d'ordre 1 si $id_S \in L^1(S, P_X)$, la sommabilité et l'espérance ne dépendent donc que de la loi P_X . Alors $E[|X|] = \sum_{s \in S} |id_S(s)| P(X = s) = \|id_S\|_{L^1(S, P_X)}$. Si Ω est lui même dénombrable, alors toutes les variables sont discrètes et $L^1(\Omega, P)$ est l'ensemble des variables discrètes d'ordre 1.

Remarquons que si X est bornée (et, en particulier, si $S = X(\Omega)$ est fini), la v.a. X est d'ordre 1. On note $L_d^1(\Omega)$ l'ensemble des v.a.d. d'ordre 1. C'est un espace vectoriel (cf plus bas).

Exemple 39. Si X est la v.a. de l'exemple 36, on a $E[X] = 1\frac{3}{5} + 2\frac{4}{15} + 3\frac{1}{10} + 4\frac{1}{35} + 5\frac{1}{210} = \frac{11}{7}$.

Exemple 40. Si S_n est la v.a. de l'exemple 37, on a $E[S_n] = np$.

La moyenne est une forme linéaire croissante sur l'espace vectoriel $L_d^1(\Omega)$ des v.a.d. d'ordre 1. En d'autres termes, on a

Proposition 51. *Si X et Y sont deux v.a.d. d'ordre 1, alors*

1. *Pour tout couple (a, b) de scalaires, $aX + bY$ est d'ordre 1 et $E(aX + bY) = aE(X) + bE(Y)$.*
2. *$X \leq Y \Rightarrow E[X] \leq E[Y]$. En particulier, $X \geq 0 \Rightarrow E[X] \geq 0$. De plus, on a égalité seulement si $X = 0$.*
3. *Z v.a.d. et $|Z| \leq |X|$ implique Z est d'ordre 1.*
4. *$|E(X)| \leq E(|X|)$.*
5. *Pour des scalaires a, b, c , on a $E[c] = c$, $a \leq X \leq b \Rightarrow a \leq E[X] \leq b$.*
6. *Si X et Y sont indépendantes, XY est d'ordre 1 et $E(XY) = E(X)E(Y)$.*

Exercice 2. Généraliser la formule du 6 pour n variables indépendantes : Si X_1, \dots, X_n sont d'ordre 1 et indépendantes, $X_1 \dots X_n$ est d'ordre 1 et $E(X_1 \dots X_n) = E(X_1) \dots E(X_n)$. (indication : faire une récurrence en montrant simultanément que $X_1 \dots X_{k-1}, X_k, \dots, X_n$ sont indépendantes.)

Démonstration. Première Méthode utilisant la fin du chapitre 2 (pour 1,3,4,6) On réalise des variables de même loi sur un espace dénombrable. On prend $\Omega' = (X, Y)(\Omega)$ qui est au plus dénombrable comme partie d'un produit d'ensembles au plus dénombrable. On munit Ω' de la probabilité P' déterminée par $P'(\{x, y\}) = P(X = x, Y = y)$ ce qu'on appellera loi du couple. Soit $X'(x, y) = x, Y'(x, y) = y$ comme v.a.d sur Ω' . Alors par définition $P'(X' = x) = \sum_{y \in Y(\Omega)} P(X = x; Y = y) = P(X = x)$ donc X', X ont même loi donc même espérance et donc $X' \in L^1(\Omega', P')$ car X' d'ordre 1 et X', X ont même espérance. De même pour Y' et Y . (en fait on dira que le couple a même loi donc les fonctions comme $aX + bY, XY$ ont même loi que respectivement $aX' + bY', X'Y'$)

1/Comme $L^1(\Omega', P')$ est un e.v. par le théorème 41, on a $aX' + bY' \in L^1(\Omega', P')$ donc $aX + bY$ aussi. Par linéarité de la somme de séries on obtient

$$\begin{aligned} E(aX + bY) &= \sum_{(x,y) \in \Omega'} (ax + by) P'((X', Y') = (x, y)) \\ &= a \sum_{(x,y) \in \Omega'} x P'((X', Y') = (x, y)) + b \sum_{(x,y) \in \Omega'} y P'((X', Y') = (x, y)) = aE(X) + bE(Y) \end{aligned}$$

3/ on fait de même avec (X, Z) et on obtient Ω'' . Z' (donc Z) d'ordre 1 car par l'inégalité $|Z| \leq |X|$ on a $P'((X', Z') = (x, z)) = 0$ si $|x| < |z|$ donc :

$$\sum_{(x,z) \in \Omega''} |z| P'((X', Z') = (x, z)) = \sum_{(x,z) \in \Omega'' : |x| \leq |z|} |z| P'((X', Z') = (x, z)) \leq \sum_{(x,z) \in \Omega''} |x| P'((X', Z') = (x, z)) < \infty$$

4/Par la proposition 37 et les définitions :

$$|E(X)| = \left| \sum_{(x,z) \in \Omega''} xP'((X', Z') = (x, z)) \right| \leq \sum_{(x,z) \in \Omega''} |x|P'((X', Z') = (x, z)) = E(|X|)$$

6/ Comme ci dessus il faut voir $X'Y'$ d'ordre 1, mais

$$\begin{aligned} E(|X'Y'|) &= \sum_{(x,y) \in \Omega'} |xy|P'((X', Y') = (x, y)) \\ &= \sum_{(x,y) \in \Omega'} |xy|P'(X' = x)P(Y' = y) \leq \sum_{x \in X(\Omega), y \in Y(\Omega)} |xy|P'(X' = x)P(Y' = y) = E(|X|)E(|Y|) < \infty \end{aligned}$$

où l'indépendance est utilisée au début de la deuxième ligne et les définitions des sommes ensuite. Noter qu'en fait dans le cas indépendant $\Omega' = X(\Omega) \times Y(\Omega)$ et que la première inégalité est une égalité.

Ensuite on calcule de même,

$$\begin{aligned} E(X'Y') &= \sum_{(x,y) \in \Omega'} xyP'((X', Y') = (x, y)) = \sum_{(x,y) \in \Omega'} xyP'(X' = x)P(Y' = y) \\ &= \sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} xyP'(X' = x)P(Y' = y) = \sum_{x \in X(\Omega)} xP'(X' = x)E(Y') = E(X')E(Y'). \end{aligned}$$

Autre Méthode directe : 1/ L'image de $aX + bY$ est l'image de l'ensemble dénombrable $(X \times Y)(\Omega)$ par l'application $(x, y) \mapsto ax + by$ donc l'image est au plus dénombrable. Pour vérifier qu'il s'agit d'une variable aléatoire on doit voir que $\{aX + bY = k\}$ est bien dans la tribu. Mais il s'écrit comme l'union dénombrable $\bigcup_{(x,y) \in (X \times Y)(\Omega): ax+by=k} \{X = x\} \cap \{Y = y\}$ qui est bien dans \mathcal{T} car $\{X = x\}, \{Y = y\}$ le sont et la tribu est stable par unions et intersections au plus dénombrables. Enfin, la description ci-dessus est une union disjointe donc on peut calculer :

$$\begin{aligned} E(|aX + bY|) &= \sum_{k \in (aX + bY)(\Omega)} |k|P(aX + bY = k) \\ &= \sum_{k \in (aX + bY)(\Omega)} |k| \sum_{(x,y) \in X(\Omega) \times Y(\Omega): ax+by=k} P(X = x, Y = y) \\ &= \sum_{x \in X(\Omega), y \in Y(\Omega)} |ax + by|P(X = x, Y = y) \end{aligned}$$

où la dernière égalité vient d'une sommation par paquets.

En appliquant l'inégalité triangulaire on obtient

$$\begin{aligned} E(|aX + bY|) &\leq \sum_{x \in X(\Omega), y \in Y(\Omega)} (|a||x| + |b||y|)P(X = x, Y = y) \\ &= |a| \sum_{x \in X(\Omega), y \in Y(\Omega)} |x|P(X = x, Y = y) + |b| \sum_{x \in X(\Omega), y \in Y(\Omega)} |y|P(X = x, Y = y) \\ &= |a| \sum_{x \in X(\Omega)} |x|P(X = x) + |b| \sum_{y \in Y(\Omega)} |y|P(Y = y) = |a|E(|X|) + |b|E(|Y|) < \infty \end{aligned}$$

ce qui montre que $aX + bY$ est une v.a. d'ordre 1.

On fait ensuite le même calcul pour les familles sommables avec signe pour obtenir par sommation par paquets :

$$E(aX + bY) = \sum_{x \in X(\Omega), y \in Y(\Omega)} (ax + by)P(X = x, Y = y)$$

puis

$$\begin{aligned} E(aX + bY) &= a \sum_{x \in X(\Omega), y \in Y(\Omega)} xP(X = x, Y = y) + b \sum_{x \in X(\Omega), y \in Y(\Omega)} yP(X = x, Y = y) \\ &= a \sum_{x \in X(\Omega)} xP(X = x) + b \sum_{y \in Y(\Omega)} yP(Y = y) = aE(X) + bE(Y). \end{aligned}$$

2/ On montre d'abord que si $X \geq 0 \Rightarrow E[X] \geq 0$.

En effet, la famille sommable $\sum_{x \in X(\Omega)} xP(X = x)$ est à termes positifs donc de somme $E[X]$ positive. Si $X \leq Y$ alors $Y - X \geq 0$ d'où par le 1/ et le cas précédent : $E(Y) - E(X) = E(Y - X) \geq 0$.

De plus si $X \geq 0$ et $X \neq 0$ il existe k tel que $P(X = k) > 0$ d'où $E(X) \geq kP(X = k) > 0$. On obtient le cas d'égalité par contraposée.

4/ On remarque qu'en utilisant une union disjointe à 2 termes, puis en sommant par paquets :

$$\begin{aligned} E(|X|) &= \sum_{k \in |X(\Omega)|} kP(|X| = k) = \sum_{k \in |X(\Omega)|} k(P(X = k) + P(X = -k)) \\ &= \sum_{(k, \epsilon) \in |X(\Omega)| \times \{+1, -1\}} |\epsilon k|P(X = \epsilon k) \\ &= \sum_{x \in X(\Omega)} |x|P(X = x) \end{aligned}$$

où la dernière égalité vient en remarquant que $X(\Omega) \subset \{\epsilon k : (k, \epsilon) \in |X(\Omega)| \times \{+1, -1\}\}$ et que les probabilités $P(X = \epsilon k)$ sont non nulles juste sur ce sous-ensemble. Alors comme pour une famille sommable

$$\sum_{x \in X(\Omega)} |x|P(X = x) \geq \left| \sum_{x \in X(\Omega)} xP(X = x) \right|,$$

on obtient le résultat.

3/ On part de $E(|X|) = \sum_{x \in X(\Omega)} |x|P(X = x)$ et on essaye de dominer la série définissant $E(|Z|)$. ON décompose $\{X = x\} = \cup_{z \in Z(\Omega)} \{X = x, Z = z\}$ et on peut se restreindre sur les $z \leq x$ vu l'inégalité $Z \leq X$ implique que $P(X = x, Z = z) = 0$ sinon .

Alors par union dénombrable puis l'inégalité qu'on vient d'expliquer :

$$\begin{aligned} \sum_{x \in X(\Omega)} |x|P(X = x) &= \sum_{x \in X(\Omega)} \sum_{z \in Z(\Omega)} |x|P(X = x, Z = z) \\ &= \sum_{x \in X(\Omega)} \sum_{z \in Z(\Omega)} |x|P(X = x, Z = z) \geq \sum_{x \in X(\Omega)} \sum_{z \in Z(\Omega)} |z|P(X = x, Z = z) \end{aligned}$$

puis par Fubini-Tonelli (puis union dénombrable) :

$$\sum_{x \in X(\Omega)} |x|P(X = x) \geq \sum_{z \in Z(\Omega)} \sum_{x \in X(\Omega)} |z|P(X = x, Z = z) = \sum_{z \in Z(\Omega)} |z|P(Z = z).$$

La finitude de la dernière somme donne Z d'ordre 1.

5/ est évident à partir de 2 vu $P(c = c) = 1$ donc $E(c) = cP(c = c) = c$.

6/ Si X et Y sont indépendantes, XY est à valeur l'image par le produit de l'ensemble au plus dénombrable $(X \times Y)(\Omega)$ donc à valeur au plus dénombrable. Comme pour la somme on a l'union disjointe

$$\{XY = k\} = \bigcup_{(x,y) \in (X \times Y)(\Omega): xy=k} \{X = x\} \cap \{Y = y\}$$

ce qui appartient à la tribu \mathcal{T} .

Pour voir que XY est d'ordre 1, on regarde la somme

$$\begin{aligned} \sum_{k \in XY(\Omega)} |k|P(XY = k) &= \sum_{k \in XY(\Omega)} |k| \sum_{(x,y) \in (X \times Y)(\Omega): xy=k} P(X = x, Y = y) \\ &= \sum_{k \in XY(\Omega)} \sum_{(x,y) \in (X \times Y)(\Omega): xy=k} |x||y|P(X = x)P(Y = y) \end{aligned}$$

et la dernière égalité est obtenue par indépendance. En sommant par paquets, on obtient :

$$\sum_{k \in XY(\Omega)} |k|P(XY = k) = \sum_{(x,y) \in (X \times Y)(\Omega)} |x||y|P(X = x)P(Y = y) = E(|X|)E(|Y|) < \infty$$

d'où le produit XY est bien d'ordre 1.

Enfin, en appliquant le même calcul aux familles sommables signées on obtient l'égalité voulue $E(XY) = E(X)E(Y)$ (cf cours d'amphi). \square

On peut aussi s'intéresser au calcul de la moyenne d'une fonction quelconque d'une v.a.d. en généralisant la relation précédente. Ceci permettra en même temps d'obtenir différentes caractérisations de la loi d'une v.a.d. à travers le calcul de la moyenne d'une fonction bien choisie. Ceci nous conduit au théorème de transfert.

4.5 Théorème de transfert

Théorème 52. Soit $X : \Omega \rightarrow \mathbb{R}$ une v.a.d., $S = X(\Omega)$ et $h : S \rightarrow \mathcal{C}$ une fonction telle que $\sum_{s \in S} |h(s)|P(X = s)$ soit convergente. Alors, la v.a.d. $h(X)$ est d'ordre 1 et on a :

$$E[h(X)] = \sum_{s \in S} h(s)P(X = s).$$

Démonstration. $h(S)$ est dénombrable comme image par une application d'un ensemble dénombrable (si $f : \mathbb{N} \rightarrow S$ est une surjection, $h \circ f : \mathbb{N} \rightarrow h(S)$ est une surjection). On a l'union disjointe

$$\{h(X) = y\} = \bigcup_{x \in X(\Omega): h(x)=y} \{X = x\}$$

qui montre que $h(X)$ est bien une v.a. discrète.

Donc en prenant des unions dénombrables disjointes puis sommation par paquets :

$$\sum_{y \in h(S)} |y|P(h(X) = y) = \sum_{y \in h(S)} \sum_{x \in X(\Omega): h(x)=y} |h(x)|P(X = x) = \sum_{s \in S} |h(s)|P(X = s) < \infty.$$

Cela montre que $h(X)$ est d'ordre 1.

Pour voir l'égalité on écrit de même en utilisant la sommation par paquets des familles sommables :

$$E(h(X)) = \sum_{y \in h(S)} yP(h(X) = y) = \sum_{y \in h(S)} \sum_{x \in X(\Omega):h(x)=y} h(x)P(X = x) = \sum_{s \in S} h(s)P(X = s).$$

□

Remarquons que lorsque X et Y sont égales en loi, on voit que, pour toute fonction h , $E[h(X)] = E[h(Y)]$, pourvu que l'un des deux membres existe. En particulier, X et Y ont mêmes paramètres (moyenne, variance, f.r., f.c. etc).

Proposition 53. X et Y v.a.d. sont indépendantes si et seulement si

$$E[h(X)g(Y)] = E[h(X)]E[g(Y)],$$

pour toutes fonctions h, g telle que $h(X), g(Y)$ d'ordre 1 tel que $h(X)g(Y)$ d'ordre 1. De plus, pour caractériser l'indépendance il suffit de prendre h, g bornées.

Démonstration. La condition suffisante est évidente car en prenant $h = 1_A, g = 1_B$ des indicatrices (qui sont bornées) on retrouve la définition de l'indépendance.

Réciproquement, il suffit de voir que $h(X), g(Y)$ sont indépendantes par la proposition 51. 6. Mais on a vu à la proposition précédente que les événements $\{h(X) = z\} = \cup_{x \in X(\Omega):h(x)=z}\{X = x\}$ et de même $\{g(Y) = t\} = \cup_{y \in X(\Omega):h(y)=t}\{Y = y\}$ sont donc indépendants comme unions disjointes d'évènements indépendants (lemme 48).

□

4.6 Moments d'une v.a.d.

Définition 56. Soit k un entier positif. La v.a.d. X est dite d'ordre k si $\sum_{s \in S} |s|^k p_s$ est convergente.

On voit que si X est d'ordre k , alors elle est d'ordre j pour tout $j \leq k$. (cf fin du chapitre 2). Remarquons que si X est bornée (et, en particulier, si $S = X(\Omega)$ est fini), la v.a. X est d'ordre k pour tout k . On note $L_d^k(\Omega)$ l'ensemble des v.a.d. d'ordre k . En particulier, dans le cas le plus fréquent pour nous, si Ω dénombrable on a $L_d^k(\Omega) = L^k(\Omega, P)$ selon la notation du chapitre 2.

C'est un espace vectoriel. La formule ci-dessous vient aussi par transfert :

Définition 57. Soit X une v.a. d'ordre k . Alors Le *moment d'ordre k* de X est défini par la quantité

$$m_k = E[X^k] = \sum_{s \in S} s^k p_s.$$

On suppose que X est une v.a.d. d'ordre 2 de moyenne m .

Définition 58. La *variance* de X est définie par la quantité $v = \sigma^2 = V[X] = E[(X - m)^2] = E[X^2] - m^2$. L'*écart-type* de X est la quantité $\sigma = (V[X])^{1/2}$.

5 Couples aléatoires (avec de nombreux résultats admis)

On considère la tribu borélienne $\mathcal{B}(\mathbb{R}^2)$ (engendrée par les produits d'intervalles $[a, b] \times [c, d]$ qui contient donc les unions dénombrables de tels produits).

Comme pour les variables aléatoires, on va définir la loi du couple comme probabilité, ayant une fonction de répartition, une fonction caractéristique etc. permettant de la décrire. On verra ensuite le cas particulier des v.a. discrètes et continues.

Définition 59. (facultatif) La *loi du couple* de v.a. $(X, Y) : \Omega \rightarrow \mathbb{R}^2$ est la probabilité $P_{(X,Y)}$ sur $\mathcal{B}(\mathbb{R}^2)$ défini par :

$$P_{(X,Y)}(B) = P((X, Y) \in B).$$

Elle est caractérisée par le cas $B = [a, b] \times [c, d]$

$$P_{(X,Y)}([a, b] \times [c, d]) = P(X \in [a, b], Y \in [c, d]).$$

Les lois P_X et P_Y sont dites lois marginales et sont déterminées par :

$$P_X(A) = P_{(X,Y)}(A \times \mathbb{R}), P_Y(B) = P_{(X,Y)}(\mathbb{R} \times B).$$

Définition 60. La fonction de répartition de $(X, Y) : \Omega \rightarrow \mathbb{R}^2$ est la fonction $F_{(X,Y)} : \mathbb{R}^2 \rightarrow [0, 1]$ définie par

$$F_{(X,Y)}(x, y) = P(X \leq x, Y \leq y).$$

Proposition 54. $x \mapsto F_{(X,Y)}(x, y)$ et $y \mapsto F_{(X,Y)}(x, y)$ sont croissantes et

$$\lim_{x \rightarrow \infty} F_{(X,Y)}(x, y) = F_Y(y), \lim_{y \rightarrow \infty} F_{(X,Y)}(x, y) = F_X(x).$$

$F_{(X,Y)}$ caractérise $P_{(X,Y)}$.

Exemple 41. Soit (X, Y) prenant les valeurs $(1, 1), (1, 2), (2, 1)$ avec les probabilités $\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$. On a donc

$$P(X = 1) = P(X = 1, Y = 1) + P(X = 1, Y = 2) = \frac{3}{4},$$

et

$$P(Y = 1) = P(X = 1, Y = 1) + P(X = 2, Y = 1) = \frac{3}{4},$$

Or

$$P(X = 1, Y = 1) = \frac{1}{2} \neq \left(\frac{3}{4}\right)^2 = P(X = 1)P(Y = 1)$$

donc X et Y ne sont pas indépendantes. On a surtout introduit les couples pour étudier le cas des v.a. non indépendantes. Plus généralement la loi marginale est donnée par $P(X = 1) = \frac{3}{4}$

On résume souvent cela dans le tableau :

	1	2	Y
1	0.5	0.25	0.75
2	0.25	0	0.25
X	0.75	0.25	1

Le tableau suivant avec les mêmes marginales correspond à des variables indépendantes. On voit au passage que les marginales ne suffisent pas à caractériser la loi du couple (le tableau contient plus d'informations que les dernières lignes et colonnes).

	1	2	Y'
1	0.5625	0.1875	0.75
2	0.1875	0.0625	0.25
X'	0.75	0.25	1

Définition 61. Soient $X, Y \in L^1(\Omega)$ des variables d'ordre 1, l'espérance de (X, Y) est le vecteur $(E(X), E(Y))$.

Soient $X, Y \in L^2(\Omega)$ des variables d'ordre 2, la *matrice de covariance* de (X, Y) est la matrice 2×2 (symétrique) :

$$\begin{pmatrix} E((X - E(X))^2) & E((X - E(X))(Y - E(Y))) \\ E((X - E(X))(Y - E(Y))) & E((Y - E(Y))^2) \end{pmatrix}.$$

On note $Cov(X, Y) = E((X - E(X))(Y - E(Y)))$.

5.1 Couples de v.a. discrètes

Si X, Y sont des v.a. discrètes, le couple (X, Y) est à valeur $(X, Y)(\Omega) \subset X(\Omega) \times Y(\Omega)$ qui est encore fini ou dénombrable.

La loi est donc caractérisée par les valeurs :

$$P(X = i, Y = j), (i, j) \in (X, Y)(\Omega).$$

Elles vérifient bien sûr

$$\sum_{(i,j) \in (X,Y)(\Omega)} P(X = i, Y = j) = 1.$$

Un Théorème de transfert énonce alors que pour toute fonction $h : (X, Y)(\Omega) \rightarrow \mathbb{R}$, on a :

$$E(h(X, Y)) = \sum_{(i,j) \in (X,Y)(\Omega)} h(i, j)P(X = i, Y = j).$$

Cela permet de trouver tous les paramètres du couple.

Exemple 42. Une urne contient r boules rouges et v boules vertes, soit $b = r + v$ boules au total. On tire de façon successive avec remise n boules de l'urne. On note X le nombre de boules rouges tirées, Y le nombre de boules vertes tirées. Calculons la loi du couple (X, Y) . On a $(X, Y)(\Omega) = \{(k, l) \in \mathbb{N}^{*2}, k + l = n\}$. On calcule $\Omega = \llbracket 1, b \rrbracket^2$, $Card(\Omega) = b^n$, $Card(X = k, Y = l) = r^k v^l \binom{k}{l}$ (choix des boules rouges tirées r^k , des vertes v^l et des position des rouges et des vertes $\binom{k}{l}$) donc la loi de (X, Y) est déterminée par :

$$P(X = k, Y = l) = \frac{k!}{l!(k-l)!} \left(\frac{r}{b}\right)^k \left(\frac{v}{b}\right)^l.$$

La loi marginale de X est donnée par $P(X = k) = P(X = k, Y = n - k)$, c'est donc la loi binomiale $\mathcal{B}(n, \frac{r}{b})$, on a $Y = n - X$.

6 Compléments sur les variables aléatoires discrètes

En plus de la f.r., les deux fonctions suivantes constituent des outils intéressants pour caractériser la loi de X .

6.1 Variance et Écart-type d'une v.a.d. Variable centrée réduite

On suppose que X est une v.a.d. d'ordre 2 de moyenne m .

Définition 62. La *variance* de X est définie par la quantité $v = \sigma^2 = V[X] = E[(X - m)^2] = E[X^2] - m^2$. L'*écart-type* de X est la quantité $\sigma = (V[X])^{1/2}$.

La variable centrée réduite associée à X est $U = \frac{X-m}{\sigma}$. On a $E[U] = 0$ et $V[U] = 1$.

Remarque 21. 1. La moyenne et l'écart-type d'une v.a. X constituent un résumé intéressant de la répartition de X . La moyenne donne une idée de l'ordre de grandeur de la variable et l'écart-type une idée de la dispersion des valeurs de X autour de leur moyenne. Ces deux paramètres interviennent dans plusieurs inégalités.

2. Le recours à la variable centrée réduite permet et facilite certains calculs. Il permet aussi de comparer entre elles de façon cohérente plusieurs variables.

La variance est définie sur l'espace vectoriel $L^2_d(\Omega)$ des v.a.d. d'ordre 2.

Proposition 55. *si X et Y sont deux v.a.d. d'ordre 2, a, b, c des constantes réelles, on a les propriétés suivantes :*

1. $V[c] = 0$, $V[X] \geq 0$ et $V[X] = 0 \Leftrightarrow X = m$ pour une constante $m = E(X)$.
2. $V[X] = E[X^2] - (E[X])^2$.
3. $V(aX + b) = a^2V(X)$.
4. Si X et Y sont indépendantes, $V(X + Y) = V(X) + V(Y)$.

La réciproque du dernier résultat est FAUSSE.

Démonstration. 1/ Comme $E(c) = c$ pour une constante $(c - E(c))^2 = 0$ d'où $V(c) = 0$. Pour voir l'implication si $V[X] = 0$. Il suffit d'utiliser le cas d'égalité de la proposition 51.2 vu que $E((X - m)^2) = 0$, $(X - m)^2 \geq 0$, on obtient $(X - m)^2 = 0$.

2/ En développant $(X - m)^2 = X^2 - 2mX + m^2$ et en utilisant la linéarité de l'espérance, (vu X^2, mX d'ordres 1) : $V[X] = E(X^2) + m^2 - E(2mX) = E(X^2) + m^2 - 2m^2 = E[X^2] - (E[X])^2$.

3/ Comme $E(aX + b) = aE(X) + b$ on obtient $aX + b - E(aX + b) = a(X - E(X))$ d'où le résultat en prenant un carré et par linéarité de l'espérance.

4/ En développant, on obtient $V(X + Y) = V(X) + V(Y) + 2E(XY) - 2E(X)E(Y)$. Dans le cas indépendant on a vu $E5XY) = E(X)E(Y)$ d'où le résultat. \square

6.2 Inégalités de Markov et de Bienaymé-Tchébychev

Proposition 56 (Inégalité de Markov). . *Soit Z une v.a.d positive et $a > 0$. Alors,*

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

Démonstration. Il suffit de noter en séparant la somme et appliquant des inégalités termes à termes par positivité des $b \in Z(\Omega)$:

$$\begin{aligned} E(Z) &= \sum_{b \in Z(\Omega)} bP(Z = b) = \sum_{b \in Z(\Omega): b \geq a} bP(Z = b) + \sum_{b \in Z(\Omega): b < a} bP(Z = b) \\ &\geq \sum_{b \in Z(\Omega): b \geq a} aP(Z = b) + \sum_{b \in Z(\Omega): b < a} 0P(Z = b) = aP(Z \geq a). \end{aligned}$$

□

On obtient la conséquence immédiate vu $P(|X - m| \geq \epsilon) = P((X - m)^2 \geq \epsilon^2)$:

Proposition 57 (Inégalité de Bienaymé-Tchébychev). *Soit X une v.a. de moyenne m et d'écart-type σ . Alors, pour tout $\epsilon > 0$,*

$$P(|X - m| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

Exemple 43. Soit X le nombre de rames de métro qui passent entre 16H et 17H en une station donnée. On a observé que le nombre moyen de rames est $E(X) = 55$ et que la variance est $V(X) = 20$. La probabilité que le nombre de rames soit compris entre 50 et 60 est au moins $P(X \in [50, 60]) = P(|X - E(X)| \leq 5) \geq 1 - P(|X - E(X)| \geq 5) \geq 1 - \frac{V(X)}{5^2} = 1 - 20/25 = 1/5$.

Voici une conséquence importante, la loi faible des grands nombres, que l'on renforcera au prochain chapitre.

Théorème 58. (*Loi faible des grands nombres*) *Soit $(X_n)_{n \geq 1}$ une suite de variables indépendantes de même loi, telle que $X_n \in L^2(\Omega)$ est d'ordre 2 et $m = E(X_n)$. Alors pour tout $\epsilon > 0$*

$$P\left(\left|\frac{X_1 + \dots + X_n}{n} - m\right| > \epsilon\right) \rightarrow_{n \rightarrow \infty} 0.$$

Démonstration. En effet en appliquant l'inégalité de Tchebychev à $Y_n = \frac{X_1 + \dots + X_n}{n}$ qui vérifie $E(Y_n) = m$, $V(Y_n) = \frac{1}{n^2}(\sum_{i=1}^n V(X_i)) = \frac{V(X_1)}{n}$ par indépendance et équidistribution, on obtient :

$$P\left(\left|\frac{X_1 + \dots + X_n}{n} - m\right| > \epsilon\right) \leq \frac{V(Y_n)}{\epsilon^2} = \frac{V(X_1)}{n\epsilon^2} \rightarrow 0.$$

□

6.3 Fonctions génératrice d'une v.a.d.

Soit $X : \Omega \rightarrow \mathbb{R}$ une v.a.d. et $S = X(\Omega)$. Les formules sont des applications directes du théorème de transfert.

Définition 63. La *fonction génératrice* (f.g.) de X est définie par

$$g_X(t) = E(t^X) = \sum_{s \in S} t^s p_s,$$

pour tout $t \in \mathbb{R}$ tel que la somme est finie. Le plus souvent g_X est définie lorsque $S \subset \mathbb{N}$, c'est alors une série entière.

On remarquera $g_X(t)$ est toujours définie pour $|t| \leq 1$ car la série est sommable par produit d'une famille sommable et d'une famille bornée. On verra au chapitre suivant que g_X caractérise la loi. De même, on a :

Définition 64. La *fonction génératrice* (f.g.) de (X, Y) est définie par

$$g_{(X, Y)}(t, s) = E[t^X s^Y],$$

pour tout $t, s \in [0, 1]^2$. $g_{(X, Y)}$ caractérise $P_{(X, Y)}$.

Proposition 59. *Les assertions suivantes sont équivalentes :*

1. X et Y sont indépendantes
- 2.

$$\forall (t, s) \in [0, 1]^2, g_{(X, Y)}(t, s) = g_X(t)g_Y(s),$$

Démonstration. Le sens direct vient de la formule de la proposition (53). La réciproque vient de ce que la fonction génératrice caractérise la loi (ce qu'on verra au chapitre suivant.) \square

En considérant les fonctions puissances, le théorème de transfert nous conduit à considérer les différentes notions suivantes.

Chapitre 7

Séries génératrices

On a déjà introduit la fonction génératrice d'un v.a. discrète :

$$g_X(t) = E(t^X) = \sum_{s \in X(\Omega)} t^s p_s,$$

Plus généralement la série génératrice d'une suite (b_n) est la série entière $g_n(x) = \sum_{n=1}^{\infty} b_n x^n$. On l'utilise en dénombrement pour calculer b_n et en probabilité pour aider à différents calculs. Ce chapitre ne sera pas une étude systématique de toutes les techniques des séries génératrices. Mais on présente quelques méthodes pour exploiter la variable fonctionnelle x pour obtenir des équations aidant à faire des calculs.

On commence par des calculs utilisant la dérivation par rapport à x . On continue avec des calculs plus algébriques. On vérifie des équations algébriques en exploitant le théorème de Fubini pour les séries.

1 Dérivation de la série génératrice

Exemple 44. (Cas polynomiale) Calculons $\sum_{k_1+k_2+k_3=n, k_i \geq 0} k_1 \frac{n!}{k_1! \cdots k_n!}$. Comme par la formule du multinôme

$$\sum_{r_1+\dots+r_n=p, r_i \geq 0} \frac{p!}{r_1! \cdots r_n!} x^{r_1} = (x+n-1)^p.$$

En dérivant par rapport à x on obtient :

$$\sum_{r_1+\dots+r_n=p, r_i \geq 0} \frac{p!}{r_1! \cdots r_n!} r_1 x^{r_1-1} = p(x+n-1)^{p-1}.$$

Donc en évaluant en 1 on obtient :

$$\sum_{k_1+k_2+k_3=n, k_i \geq 0}^n k_1 \frac{n!}{k_1! k_2! k_3!} = n3^{n-1}.$$

En général, on rappelle le théorème général (cf analyse 3) permettant de dériver une série entière.

Proposition 60. Soit $\sum a_n x^n$ une série entière de la variable $x \in \mathbb{R}$, de rayon de convergence $R > 0$ et f sa somme définie dans l'intervalle de convergence $I =]-R, R[$ par $f(x) = \sum_{n=0}^{\infty} a_n x^n$. Alors

f est de classe C^∞ sur I et ses dérivées s'obtiennent par dérivation terme à terme de la série. Ainsi, pour tout $k \in \mathbb{N}^*$ et tout $x \in D$,

$$f^{(k)}(x) = \sum_{n=k}^{\infty} \frac{n!}{(n-k)!} a_n x^{n-k}.$$

En particulier, $\forall k \in \mathbb{N}$, $f^{(k)}(0) = k!a_k$. De plus $f^{(k)}$ est somme de la série entière $\sum \frac{(n+k)!}{n!} a_{n+k} x^n$ de rayon de convergence R .

Proposition 61. Pour $X : \Omega \rightarrow \mathbb{N}$ variable entière, g_X caractérise la loi de X .

Démonstration. En effet, comme $g_X(t) = \sum_{k=0}^{\infty} t^k P(X = k)$ la proposition donne $g_X^{(k)}(0) = k!P(X = k)$. Donc g_X détermine tous les $P(X = k)$ donc la loi. \square

Exemple 45. Pour une variable aléatoire entière X , on a

$$g'_X(t) = \sum_{k \in \mathbb{N}} t^{k-1} k p_k$$

pour $t < 1$. En particulier, si X d'ordre 1, on a aussi la dérivée à gauche

$$g'_X(1) = g'_X(1^-) = E(X).$$

En effet par le théorème de L1 de limite de la dérivée. il suffit de voir $\lim_{t \rightarrow 1^-} g'_X(t) = E(X)$. Mais si la somme $\sum_{k \in \mathbb{N}} k p_k$ en dehors d'un ensemble I $\sum_{k \notin I} t^{k-1} k p_k \leq \sum_{k \notin I} k p_k \leq \epsilon$ et la somme finie $\sum_{k \in I} t^{k-1} k p_k \rightarrow \sum_{k \in I} k p_k$ donc un argument à 3ϵ donne $\sum_{k \in \mathbb{N}} t^{k-1} k p_k \rightarrow \sum_{k \in \mathbb{N}} k p_k$.

Exemple 46. Calculons ainsi la moyenne d'une variable géométrique $\mathcal{G}(\mathbb{N})$.

$g_X(t) = \sum_{k=1}^{\infty} p q^{k-1} t^k = \frac{pt}{1-qt}$ donc $(g_X)'(t) = \frac{p(1-qt)+qpt}{(1-qt)^2} = \frac{p}{(1-qt)^2}$ donc en prenant la limite $t \rightarrow 1$ on obtient :

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

Exemple 47. Pour une variable aléatoire entière X , si g_X rayon de convergence $R > 1$. On a aussi en appliquant le théorème et la formule de transfert : $g_X^{(k)}(1) = E(X(X-1)\dots(X-k+1))$.

Exemple 48. Calculons ainsi la variance d'une variable géométrique $\mathcal{G}(\mathbb{N})$.

$g_X(t) = \sum_{k=1}^{\infty} p q^{k-1} t^k = \frac{pt}{1-qt}$ donc $(g_X)''(t) = \frac{2pq}{(1-qt)^3}$ donc en prenant la limite $t \rightarrow 1$ on obtient : $E(X(X-1)) = \frac{2pq}{p^3}$ et $Var(X) = E(X(X-1)) + E(X) - E(X)^2 = \frac{2q+p-1}{p^2} = \frac{q}{p^2}$.

2 Première application en combinatoire : Nombre de solutions de l'équation $\sum_{k=1}^n x_k = p$ (facultatif, non vu en cours)

Proposition 62. Le nombre de solutions de l'équation $\sum_{k=1}^n x_k = p$ où, pour tout $k \in \llbracket 1, n \rrbracket$, $x_k \in \mathbb{N}^*$, est $\binom{p-1}{n-1}$.

Démonstration. Forcément $p \geq n$ pour avoir une solution non nulle. Soit $b_{n,p}$ le nombre de solutions. En faisant le produit de Cauchy de séries absolument convergentes (par récurrence) pour $x < 1$ (série entière à l'intérieur du rayon de convergence) :

$$\frac{x^n}{(1-x)^n} = \sum_{x_1=1}^{\infty} \dots \sum_{x_n=1}^{\infty} x^{\sum_{k=1}^n x_k} = \sum_{p=n}^{\infty} \sum_{\substack{\sum_{k=1}^n x_k=p \\ x_k \geq 1}} x^{\sum_{k=1}^n x_k} = \sum_{p=n}^{\infty} b_{n,p} x^p.$$

D'où par la formule de Taylor pour $q \geq 0$

$$b_{n,q+n} = \left(\frac{1}{q!} \left(\frac{d}{dx}\right)^q \frac{1}{(1-x)^n}\right)_{x=0} = \left(\frac{1}{q!} \left(\frac{d}{dx}\right)^{p-1} \frac{n}{(1-x)^{n+1}}\right)_{x=0} = \frac{n(n+1)\dots(n+q-1)}{q!} = \binom{n+q-1}{q} = \binom{n+q-1}{n} \quad \square$$

Exercice 3. Le nombre de solutions de l'équation $\sum_{k=1}^n x_k = p$ où, pour tout $k \in \llbracket 1, n \rrbracket$, $x_k \in \mathbb{N}$, est $K_n^p = \binom{n+p-1}{n-1}$.

Exercice 4. Le nombre de solutions de l'équation $\sum_{k=1}^n x_k = p$ où, pour tout $k \in \llbracket 1, n \rrbracket$, $x_k \geq s_k$, est $\binom{p-s+n-1}{n-1}$ avec $s = \sum_{k=1}^n s_k$.

3 Une Autre Application de la méthode des séries génératrices : nombre de parenthésages (facultatif, non vu en cours)

Exemple 49. Soit C_n le nombre de parenthésages à n parenthèses ouvrantes. La première parenthèse doit se refermer après avoir ouvert entre 0 (si elle se referme tout de suite) et $n-1$ parenthèses, d'où la relation de récurrence :

$$C_n = \sum_{k=0}^{n-1} C_k C_{n-k}, C_0 = 1.$$

Si on pose $T(x) = \sum_{k=0}^{\infty} C_k x^k$ cette série a rayon de convergence au moins $1/4$ car $C_n \leq 2^{2n}$ (le nombre de parenthésage bien ou mal formé de $2n$ est $\binom{2n}{n} \leq 2^{2n}$.)

De plus la relation devient

$$T(x) = \sum_{n=0}^{\infty} \sum_{k=0}^{n-1} C_k C_{n-k} x^k x^{n-k} = 1 + x \sum_{m=0}^{\infty} \sum_{k=0}^m C_k C_{n-k} x^k x^{m-k} = 1 + xT(x)^2$$

(en prenant $m = n-1$ et en regardant le produit de Cauchy de T et lui même).

En résolvant l'équation du second degré, on obtient :

$$\begin{aligned} T(x) &= \frac{1 - \sqrt{1 - 4x}}{2x} \\ &= -\frac{1}{2x} \sum_{n=1}^{\infty} \frac{\frac{1}{2}(-\frac{1}{2})(-\frac{3}{2})\dots(-\frac{2n+3}{2})}{n!} (-4x)^n \\ &= \frac{1}{2x} \sum_{n=0}^{\infty} \frac{(1.3\dots(2n-3))}{n!} (2x)^n \\ &= \sum_{n=1}^{\infty} \frac{(2n-2)!}{n!(n-1)!} x^{n-1}. \end{aligned}$$

Donc $C_n = \frac{1}{n+1} \binom{2n}{n}$.

3.1 Application au calcul de loi de sommes de v.a. indépendantes

On a traité en cours d'amphi l'exo suivant :

Exemple 50. Soit X de loi de Poisson $\mathcal{P}(\lambda)$ et Y de Poisson $\mathcal{P}(\mu)$ indépendantes. Calculer la fonction génératrice $g_{X+Y}(t)$ et déduire que $X + Y$ est de loi $\mathcal{P}(\lambda + \mu)$.

Chapitre 8

Théorème de Moivre-Laplace I : Cadre et Énoncé

Soit $S_n = B_1 + \dots + B_n$ une somme de variables de Bernoulli indépendantes $B(1, p)$. On rappelle que S_n suit la loi binomiale $B(n, p)$. Le but de ce chapitre est de comprendre l'écart de S_n à sa moyenne np . On verra que sur l'échelle \sqrt{n} , les variables $S_n - E(S_n)$ se comportent asymptotiquement comme une variable qui n'est pas à support dénombrable, la loi gaussienne (ou normale). On va donc d'abord brièvement présenter les variables aléatoires continues et quelques exemples, dont le principal pour nous dans la suite, la loi gaussienne. On va suivre dans un premier temps une preuve élémentaire (venant du livre *Probability* de Shiryaev). Puis on énoncera un résultat plus général en donnant une preuve facultative qui utilise une boîte noire : le théorème de Paul Lévy que vous verrez en L3.

1 Introduction aux Variables aléatoires continues (ou à densité)

Les notions relatives aux variables aléatoires continues (v.a.c.) sont essentiellement les mêmes que celles des v.a.d., seule l'apparence du mode de calcul change, le signe \sum étant remplacé par le signe \int .

Nous suivrons le plan d'étude d'une v.a.d. en apportant les modifications nécessaires et en signalant les différences éventuelles. Lorsque X est une v.a.d., $S = X(\Omega)$ est fini ou dénombrable et la loi de X est définie à partir d'une suite à termes positifs c'est-à-dire une fonction sur \mathbb{N} (ou S) positive et de somme 1. Dans le cas d'une v.a.c. X , $S = X(\Omega)$ n'est pas dénombrable et la loi de X sera définie à partir d'une fonction sur \mathbb{R} positive et d'intégrale 1 et cela nous amène à parler de densité.

1.1 Densité d'une probabilité

Une densité de probabilité f (sur \mathbb{R}) est une fonction positive $f : \mathbb{R} \rightarrow \mathbb{R}_+$, intégrable (au sens de Riemann) telle que $\int_{-\infty}^{\infty} f(x)dx = 1$. (Par exemple f continue par morceau sur tout segment et $\int_{-\infty}^{\infty} f(x)dx := \lim_{K \rightarrow \infty} \int_{-K}^K f(x)dx = 1$)

Définition 65. Une variable aléatoire $X : \Omega \rightarrow \mathbb{R}$ est dite continue (v.a.c.) si sa loi P_X admet une densité f , c'est-à-dire si l'on a pour tout intervalle $I = [a, b] \subset \mathbb{R}$, $P(X \in [a, b]) = \int_a^b f(x)dx$. Cette densité est alors unique et on dit que f est la densité de X ou que X admet la densité f .

Remarque 22. Soit X une v.a.c. de densité f

1. Pour tout $a \in \mathbb{R}$, $P(X = a) = \int_{\{a\}} f(x)dx = 0$. Ainsi, contrairement à une v.a.d., une v.a.c. prend avec la probabilité 0 toute valeur unique (singleton).
2. Soit $S = X(\Omega)$. Dans les cas usuels, S est un intervalle et, quand il est distinct de \mathbb{R} , on a $\int_{S^c} f(x)dx = P(X \in S^c) = P(\emptyset) = 0$. Ceci impose à f d'être nulle en dehors de S .

Ainsi, lorsque la v.a. est continue, on considère que sa loi est correctement définie quand on précise sa densité qui doit être nulle en dehors de l'ensemble de ses valeurs.

1.2 Lois à densité classiques

Mis à part pour la loi normale, nous vous laissons en exercice de vérifier que les fonctions énoncés sont bien des densités de probabilités.

Exemple 51. La **loi uniforme sur le segment** $[a, b]$, notée $\mathcal{U}([a, b])$ est la loi de densité $f(x) = \frac{1}{b-a}1_{[a,b]}(x)$.

Exemple 52. La **loi exponentielle de paramètre** λ , notée $\mathcal{E}(\lambda)$, a pour densité $f(x) = \lambda e^{-\lambda x}1_{[0,\infty[}(x)$. Elle modélise la durée de vie d'un matériel ou un processus sans mémoire car $P(X - t > s | X > t) = P(X > s)$. En effet, sa fonction de répartition est $F_X(x) = 1_{[0,\infty[}(x)(1 - e^{-\lambda x})$.

Exemple 53. La **loi normale** $\mathcal{N}(m, \sigma^2)$ (ou loi gaussienne) a pour densité :

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

Elle intervient dans le théorème centrale limite, dont nous allons voir un cas particulier dans la suite de ce chapitre. Il faut vérifier qu'il s'agit bien d'une densité.

On veut donc voir $\int_{-\infty}^{\infty} dx \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} = 1$

D'abord par changement de variable $y = x - m$, il suffit du cas $m = 0$, ensuite par $z = y/\sigma$ $\int_{-K}^K dy \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} = \int_{-K/\sigma}^{K/\sigma} dz \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ il suffit donc du cas $\sigma = 1$. Enfin on considère le carré et fait un changement de variable en polaire et on obtient :

$$\left(\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx \right)^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}} dx dy = \int_0^{\infty} dr r e^{-\frac{r^2}{2}} \int_0^{2\pi} d\theta = 2\pi$$

Exemple 54. La **loi semicirculaire** (ou de Wigner) a pour densité :

$$f(x) = 1_{[-R,R]}(x) \frac{2}{\pi R^2} \sqrt{R^2 - x^2}.$$

Elle intervient dans la loi des valeurs propres des grandes matrices aléatoires.

Exemple 55. La **loi de Cauchy** de paramètres l et c , notée $\mathcal{C}(l, c)$:

$$f(x) = \frac{1}{\pi} \frac{c}{c^2 + x^2}.$$

C'est la loi de $l + c\frac{X}{Y}$ avec X, Y normales standards indépendantes. Attention l est la médiane mais pas la moyenne.

Exemple 56. La **loi du χ^2 à ν degrés de liberté**, notée $\chi^2(\nu)$ a pour densité $f(x) = \frac{x^{\nu/2-1}}{2^{\nu/2}\Gamma(\nu/2)} e^{-x/2} 1_{[0,\infty[}(x)$ avec $\Gamma(r) = \int_0^{\infty} x^{r-1} e^{-x} dx$ la fonction Γ qui prolonge $\Gamma(n) = (n-1)!$. Elle intervient en statistique dans le test d'indépendance du χ^2 (et d'autres tests). Remarquez que $\chi^2(2) = \mathcal{E}(\frac{1}{2})$.

Une v.a.c. a les mêmes paramètres qu'une v.a.d. Ces paramètres s'obtiennent par une méthode de calcul appropriée comme il sera précisé dans la suite, où X désigne v.a.c. de densité f et $S = X(\Omega)$.

1.3 Théorème de transfert (admis)

Théorème 63. (admis) Soit $X : \Omega \rightarrow \mathbb{R}$ une v.a.c., $S = X(\Omega)$ et $h : S \rightarrow \mathbb{C}$ une fonction telle que $\int_{-\infty}^{\infty} |h(s)|f(s)ds$ soit convergente (avec hf Riemann intégrable). Alors, la v.a. $h(X)$ admet une moyenne et

$$E[h(X)] = \int_{-\infty}^{\infty} h(s)f(s)ds.$$

Remarquons que lorsque X et Y sont égales en loi, on voit que, pour toute fonction h , $E[h(X)] = E[h(Y)]$, pourvu que l'un des deux membres existe. En particulier, X et Y ont mêmes paramètres (moyenne, variance, f.r., f.c. etc).

En plus de la f.r., la fonction suivante constitue un outil intéressant pour caractériser la loi de X .

Définition 66. La fonction caractéristique (f.c.) de X est définie par

$$\varphi_X(t) = E[e^{itX}] = \int_{-\infty}^{\infty} e^{its} f(s)ds,$$

pour tout $t \in \mathbb{R}$.

la fonction φ_X caractérise la loi de X .

1.4 Fonction de répartition (f.r.) d'une v.a.c. (facultatif)

Définition 67. La fonction de répartition de $X : \Omega \rightarrow \mathbb{R}$ de densité f est la fonction $F_X : \mathbb{R} \rightarrow [0, 1]$ définie par

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt.$$

Quand il n'y a pas de confusion à craindre, on note plus simplement F la f.r. de X .

Proposition 64. Toute fonction de répartition F_X d'une vac est croissante, **continue** et vérifie $\lim_{x \rightarrow -\infty} F_X(x) = 0$ et $\lim_{x \rightarrow +\infty} F_X(x) = 1$.

Remarque 23. Soit X une v.a. dont la densité est inconnue mais dont les propriétés disponibles permettent de calculer sa f.r. F . Alors, lorsque F est continue et dérivable par morceaux, on peut affirmer que X est une v.a.c. qui admet la densité f obtenue en dérivant F aux points où elle est dérivable

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

et en donnant à f des valeurs arbitraires aux points où F n'est pas dérivable.

En fait, on peut montrer que la f.r. d'une v.a.c. détermine complètement la loi de cette variable.

1.5 Moyenne ou Espérance ou Moment d'ordre 1 (facultatif)

Définition 68. La v.a.c. X est dite d'ordre 1 lorsque l'intégrale $\int_{-\infty}^{\infty} |x|f(x)dx = \lim_{K \rightarrow \infty} \int_{-K}^K |x|f(x)dx < \infty$. Dans ce cas, on appelle Moyenne ou Espérance ou Moment d'ordre 1 de X la quantité

$$m = E[X] = \int_{-\infty}^{\infty} xf(x)dx = \lim_{K \rightarrow \infty} \int_{-K}^K xf(x)dx.$$

Remarquons que si X est bornée, la v.a. X est d'ordre 1. On note $L^1(\Omega)$ l'ensemble des v.a. d'ordre 1 (v.a. telles que $E(|X|) < \infty$ avec un sens non défini dans ce cours, cf L3). C'est un espace vectoriel. Attention l'ensemble des v.a.c. d'ordre 1 n'est PAS un e.v. car $0X = 0$ est discrète et pas continue.

Exemple 57. Si $X \sim \mathcal{U}([a, b])$

$$E(X) = \int_a^b \frac{x}{b-a} dx = \frac{a+b}{2}.$$

Exemple 58. Si $X \sim \mathcal{E}(\lambda)$

$$E(X) = \int_0^\infty \lambda x e^{-\lambda x} dx = \frac{1}{\lambda}.$$

Exemple 59. Si $X \sim \mathcal{N}(m, \sigma^2)$ $E(X) = m$.

La moyenne est une forme linéaire croissante sur l'espace vectoriel $L^1(\Omega)$ des v.a. d'ordre 1. En d'autres termes, on a

Proposition 65. Si X et Y sont deux v.a. d'ordre 1, alors

1. Pour tout couple (a, b) de scalaires, $aX + bY$ est d'ordre 1 et $E(aX + bY) = aE(X) + bE(Y)$.
2. $X \leq Y \Rightarrow E[X] \leq E[Y]$. En particulier, $X \geq 0 \Rightarrow E[X] \geq 0$.
3. Z v.a.d. et $|Z| \leq |X|$ implique Z est d'ordre 1.
4. $|E(X)| \leq E(|X|)$.
5. $E[c] = c$, $a \leq X \leq b \Rightarrow a \leq E[X] \leq b$.
6. Si X et Y sont indépendantes, XY est d'ordre 1 et $E(XY) = E(X)E(Y)$.

On peut aussi s'intéresser au calcul de la moyenne d'une fonction quelconque d'une v.a.d. en généralisant la relation précédente. Ceci permettra en même temps d'obtenir différentes caractérisations de la loi d'une v.a.d. à travers le calcul de la moyenne d'une fonction bien choisie. Ceci nous conduit au théorème de transfert.

1.6 Variance et Écart-type d'une v.a.c. Variable centrée réduite (facultatif)

On suppose que X est une v.a.c. d'ordre 2 (i.e. X^2 d'ordre 1) de moyenne m .

Définition 69. La variance de X est définie par la quantité $v = \sigma^2 = V[X] = E[(X - m)^2] = E[X^2] - m^2$. L'écart-type de X est la quantité $\sigma = (V[X])^{1/2}$.

La variable centrée réduite associée à X est $U = \frac{X-m}{\sigma}$. On a $E[U] = 0$ et $V[U] = 1$.

Exemple 60. La loi $\mathcal{N}(m, \sigma^2)$ a pour moyenne m et variance σ^2 , sa variable centrée réduite a pour loi $\mathcal{N}(0, 1)$ (loi normale centrée réduite).

Remarque 24. 1. La moyenne et l'écart-type d'une v.a. X constituent un résumé intéressant de la répartition de X . La moyenne donne une idée de l'ordre de grandeur de la variable et l'écart-type une idée de la dispersion des valeurs de X autour de leur moyenne. Ces deux paramètres interviennent dans plusieurs inégalités.

2. Le recours à la variable centrée réduite permet et facilite certains calculs. Il permet aussi de comparer entre elles de façon cohérente plusieurs variables.

La variance est définie sur l'espace vectoriel $L^2(\Omega)$ des v.a. d'ordre 2.

Proposition 66. *si X et Y sont deux v.a.c. d'ordre 2, on a les propriétés suivantes :*

1. $V[c] = 0$, $V[X] \geq 0$ et $V[X] = 0 \Leftrightarrow P(X = m) = 1$.
2. $V[X] = E[X^2] - (E[X])^2$.
3. $V(aX + b) = a^2V(X)$.
4. *Si X et Y sont indépendantes, $V(X + Y) = V(X) + V(Y)$.*

2 Théorème de Moivre-Laplace proprement dit

On va montrer le résultat suivant dont on verra des applications en statistique au chapitre suivant :

Théorème 67. *[Théorème de Moivre-Laplace] Soit $0 < p < 1$. Soit $(X_i)_{i \geq 1}$ une suite de v.a. de Bernoulli $B(1, p)$ indépendantes et identiquement distribuées (donc de moyenne $m = p$ et de variance $\sigma^2 = p(1 - p)$) et soit Y une v.a. de loi normale $\mathcal{N}(0, \sigma^2)$, alors*

$$P\left(\left\{\omega : \frac{X_1(\omega) + \dots + X_n(\omega) - nm}{\sqrt{n}} \in [a, b]\right\}\right) \xrightarrow{n \rightarrow \infty} P(Y \in [a, b]) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2}} dt.$$

Concrètement, le membre de gauche se calcule ainsi comme pour toute loi discrète, vu $S_n = X_1 + \dots + X_n$ de loi binomiale :

$$P_n([a, b]) := P\left(\left\{\omega : \frac{X_1(\omega) + \dots + X_n(\omega) - nm}{\sqrt{n}} \in [a, b]\right\}\right) = \sum_{k \in [a, b]} P(S_n = np + k\sqrt{n})$$

où la somme ne contient que les k tel que $np + k\sqrt{n}$ est entier pour que la probabilité indiquée soit non nulle.

3 Autres théorèmes limites (facultatif et seulement les 2 premiers énoncés vus en cours)

Les deux résultats suivants, la loi forte des grands nombres et le théorème central limite (TCL) sont très importants en statistique et en probabilité. Ils disent respectivement qu'une suite de variables aléatoires indépendantes et de même loi (i.i.d) ont une moyenne empirique qui tend presque sûrement (avec probabilité 1) vers leur moyenne commune m et que si on retranche la moyenne, l'écart à la moyenne correctement normalisé se comporte comme une loi gaussienne. Ils renforcent la loi faible des grands nombres (Théorème 58) et le Théorème de Moivre-Laplace que nous venons de voir.

Théorème 68 (Loi forte des grands nombres). *Soit $(X_i)_{i \geq 1}$ une suite de variables aléatoires d'ordre 1 indépendantes et identiquement distribuées ($P_{X_i} = P_{X_1}$), alors*

$$P\left(\left\{\omega : \frac{X_1(\omega) + \dots + X_n(\omega)}{n} \rightarrow E(X_1)\right\}\right) = 1.$$

On admet la preuve du résultat général (selon les versions elle est plus du niveau L3 ou M1) et on donne pour les curieux le cas particulier des variables d'ordre 4, mais qui est plus du niveau L3.

Preuve dans le cas $\mathbf{E}(X_1^4) < \infty$. On peut supposer $m = 0$, en remplaçant X_i , par $X_i - m$. Alors, on décompose le produit par la formule du multinôme :

$$\begin{aligned} \left(\frac{X_1 + \dots + X_n}{n}\right)^4 &= \sum_{k=1}^4 \sum_{i_1 + \dots + i_k = 4} \frac{4!}{i_1! \dots i_k!} X_1^{i_1} \dots X_n^{i_n} \\ &= \sum_{i=1}^n X_i^4 + 4 \sum_{i=1}^n \sum_{j \in [1, n] - \{i\}} X_i^3 X_j + 3 \sum_{i=1}^n \sum_{j \in [1, n] - \{i\}} X_i^2 X_j^2 \\ &\quad + 6 \sum_{i, j, k \in [1, n] \text{ distincts}} X_i X_j X_k^2 + \sum_{i, j, k, l \in [1, n] \text{ distincts}} X_i X_j X_k X_l \end{aligned}$$

donc on déduit en prenant l'espérance et utilisant l'indépendance (cf cours) :

$$\mathbf{E} \left[\left(\frac{X_1 + \dots + X_n}{n} \right)^4 \right] = \frac{E(X_1^4)}{n^3} + \frac{E(X_1^2 X_2^2) 3n(n-1)}{n^4} \leq \frac{C}{n^2}$$

donc

$$\mathbf{E} \left[\sum_{n=1}^{\infty} \left(\frac{X_1 + \dots + X_n}{n} \right)^4 \right] \leq \sum_{n=1}^{\infty} \frac{C}{n^2} < \infty$$

donc l'évènement $A = \left\{ \sum_{n=1}^{\infty} \left(\frac{X_1 + \dots + X_n}{n} \right)^4 \text{ converge} \right\}$ a probabilité $P(A) = 1$ (sans quoi l'espérance serait infinie), donc $\{\omega : \frac{X_1(\omega) + \dots + X_n(\omega)}{n} \rightarrow 0\} \supset A$ a proba 1. \square

Le théorème de Moivre-Laplace est un cas particulier du :

Théorème 69 (Théorème centrale limite TCL). *Soit $(X_i)_{i \geq 1}$ une suite de v.a. d'ordre 2 indépendantes et identiquement distribuées ($P_{X_i} = P_{X_1}$) de moyenne $m = E(X_1)$ et de variance $\sigma^2 = V(X_1)$ et soit Y une v.a. de loi normale $N(0, \sigma^2)$, alors*

$$P(\{\omega : \frac{X_1(\omega) + \dots + X_n(\omega) - nm}{\sqrt{n}} \in [a, b]\}) \xrightarrow{n \rightarrow \infty} P(Y \in [a, b]) = \int_a^b \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2}} dt.$$

3.1 Fonction caractéristiques

Définition 70. La *fonction caractéristique* (f.c.) d'une v.a.d. X est définie par

$$\varphi_X(t) = E[e^{itX}] = \sum_{s \in S} e^{its} p_s,$$

pour tout $t \in \mathbb{R}$.

3.2 Idées de preuves

On va utiliser les fonctions caractéristiques, il nous faut donc celle de la limite :

Lemme 70. Soit $X \sim \mathcal{N}(m, \sigma^2)$ de loi normale alors $\Phi_X(t) = \exp(-\frac{t^2\sigma^2}{2} + imt)$.

Démonstration. On a vu une preuve à l'exercice 8 du TD 3 de MASS 31 utilisant que la partie imaginaire est nulle par parité et le calcul de la partie réelle en établissant une équation différentielle par intégration dépendant d'un paramètre.

On donne ici une autre preuve par prolongement analytique. Par transfert, on doit montrer $\int \frac{1}{\sigma\sqrt{2\pi}} e^{ixt - \frac{(x-m)^2}{2\sigma^2}} = \exp(-\frac{t^2\sigma^2}{2} + imt)$ en faisant le changement de variables $u = (x - m)/\sigma$ on se ramène au cas $\sigma = 1, m = 0$.

En prenant $m = z$ dans le calcul de la densité, on a pour $z \in \mathbb{R}$

$$\int_{-\infty}^{\infty} dx \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2+z^2-2xz}{2}} = \int_{-\infty}^{\infty} dx \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-z)^2}{2}} = 1.$$

Pour $z \in \mathbb{C}$, en appliquant le résultat précédent

$$\sum_{n=0}^{\infty} \int_{\mathbb{R}} dx \frac{1}{\sqrt{2\pi}} \frac{|zx|^n}{n!} e^{-\frac{x^2}{2}} = \lim_{N \rightarrow \infty} \int_{\mathbb{R}} dx \frac{1}{\sqrt{2\pi}} \sum_{n=0}^N \frac{|zx|^n}{n!} e^{-\frac{x^2}{2}} \leq \int_{\mathbb{R}} dx \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2} + |zx|} \leq \exp\left(\frac{|z|^2}{2}\right) < \infty$$

La première bornitude permet d'appliquer le TCD pour les séries (ou Fubini pour la mesure discrète) et intervertir somme et série :

$$\sum_{n=0}^{\infty} z^n \int_{\mathbb{R}} dx \frac{1}{\sqrt{2\pi}} \frac{x^n}{n!} e^{-\frac{x^2}{2}} = \int_{\mathbb{R}} dx \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2} + zx}$$

la fonction de droite est donc la somme d'une série entière $\exp(\frac{z^2}{2})$ pour $z \in \mathbb{R}$, donc par identification des coefficients, elle vaut cette valeur pour tout $z \in \mathbb{C}$, en particulier pour $z = it$ et on trouve le résultat. \square

Théorème 71 (Lévy). (admis) Soit $Y_n : \Omega_n \rightarrow \mathbb{R}^d$ une suite de variables aléatoires et $Y : \Omega \rightarrow \mathbb{R}^d$ une autre v.a., alors Y_n converge en loi vers Y si et seulement si :

$$\forall t \in \mathbb{R}, \Phi_{Y_n}(t) \rightarrow_{n \rightarrow \infty} \Phi_Y(t).$$

où le théorème suivant décrit concrètement cette notion de convergence que l'on ne définira pas (on peut prendre l'un des deux théorème comme définition)

Théorème 72. (admis)

1. Soit $Y_n : \Omega_n \rightarrow \mathbb{R}^d$ une suite de variables aléatoires et μ une probabilité. Y_n converge en loi vers μ si et seulement si pour tout

$$\forall x \in \mathbb{R}, \mu(\{x\}) = 0 \Rightarrow \lim_{n \rightarrow \infty} P(Y_n \leq x) = \mu(]-\infty, x]).$$

2. (Lemme de Skohokhod)[facultatif] Soient $Y_n : \Omega_n \rightarrow \mathbb{R}^d$ une suite de vecteurs aléatoires et Y un vecteur aléatoire, $Y_n \xrightarrow{\mathcal{L}}_{n \rightarrow \infty} Y$ si et seulement si il existe un espace probabilisé (Ω, \mathcal{T}, P) , $X_n, X : \Omega \rightarrow \mathbb{R}^d$ tels que $P_{X_n} = P_{Y_n}, P_X = P_Y$ et $X_n \xrightarrow{\mathcal{L}}_{n \rightarrow \infty} X$.

Pour le cas vectoriel du théorème central limite, on a besoin de la définition d'un vecteur gaussien que l'on étudiera plus au prochain chapitre. Pour $X = (X_1, \dots, X_d)$ on note $\mathbf{Cov}(X)$ la matrice de covariance $(\mathbf{Cov}(X))_{i,j} = \text{Cov}(X_i, X_j)$. On rappelle qu'une matrice symétrique réelle est positive si ses valeurs propres sont toutes positives.

Définition 71. Soit $C \in M_d(\mathbb{R})$ une matrice symétrique positive. Un vecteur aléatoire $X : \Omega \rightarrow \mathbb{R}^d$ avec $X_i \in L^2$ est appelé **vecteur gaussien centré de covariance** C , dite loi $\mathcal{N}(0, C)$, si

$$\forall t \in \mathbb{R}^d, \Phi_X(t) = \mathbf{E}[e^{i\langle t, X \rangle}] = \exp\left(-\frac{\langle t, Ct \rangle}{2}\right).$$

On rappelle que $\langle t, s \rangle = \sum_i t_i s_i$ est le produit scalaire et donc que $\langle t, Ct \rangle = \sum_{i,j} t_i C_{ij} t_j$. On remarquera que cela coïncide avec la loi normale usuelle dans le cas $d = 1$.

On estime la fonction caractéristique des v.a. d'ordre 2.

Lemme 73. Soit un vecteur aléatoire $X : \Omega \rightarrow \mathbb{R}^d$ avec $X_i \in L^2$. Alors Φ_X est de classe \mathcal{C}^2 et son développement de Taylor en 0 est :

$$\Phi_X(t) = 1 + i \sum_{j=1}^d t_j \mathbf{E}[X_j] - \frac{1}{2} \sum_{k,j=1}^d t_k t_j \mathbf{E}[X_k X_j] + o(t^2).$$

PROOF : $\varphi : (t, \omega) \rightarrow e^{it\langle t, X(\omega) \rangle}$ est intégrable en ω et \mathcal{C}^∞ en t , et ces deux premières dérivées sont :

$$\frac{\partial \varphi}{\partial t_j} = i X_j e^{it\langle t, X(\omega) \rangle}, \quad \frac{\partial^2 \varphi}{\partial t_k \partial t_j} = -X_k X_j e^{it\langle t, X(\omega) \rangle},$$

qui sont dominées par

$$\left| \frac{\partial \varphi}{\partial t_j} \right| \leq |X_j|, \quad \left| \frac{\partial^2 \varphi}{\partial t_k \partial t_j} \right| \leq |X_k X_j|$$

qui sont intégrables car $X_i \in L^2$ et en particulier par Cauchy-Schwartz, $\mathbf{E}[|X_j X_k|] \leq \|X_j\|_2 \|X_k\|_2$. Donc d'après le théorème de dérivation avec condition de domination $\Phi_X(t) = \int \varphi(t, \omega) dP(\omega)$ est \mathcal{C}^2 et ces dérivées sont :

$$\frac{\partial \Phi_X}{\partial t_j}(t) = \mathbf{E}[i X_j e^{it\langle t, X \rangle}], \quad \frac{\partial^2 \Phi_X}{\partial t_k \partial t_j}(t) = -\mathbf{E}(X_k X_j e^{it\langle t, X \rangle}).$$

L'énoncé vient donc de la formule de Taylor-Young en 0. ■

Théorème 74. Soit (X_n) une suite de vecteurs aléatoires $X : \Omega \rightarrow \mathbb{R}^d$ indépendants de même loi et dans L^2 . Alors

$$\frac{(X_1 + \dots + X_n - n\mathbf{E}[X_1])}{\sqrt{n}} \xrightarrow{\mathcal{L}}_{n \rightarrow \infty} \mathcal{N}(0, \mathbf{Cov}(X_1)).$$

PROOF : En remplaçant X_i par $X_i - \mathbf{E}[X_i]$ on peut supposer et on suppose $\mathbf{E}[X_1] = 0$. On calcule la fonction caractéristique en utilisant d'abord l'indépendance (cf lemme ??) puis le fait que les X_i ont même loi.

$$\Phi_{\frac{X_1 + \dots + X_n}{\sqrt{n}}}(t) = \mathbf{E} \left[\exp\left(i \left\langle t, \frac{X_1 + \dots + X_n}{\sqrt{n}} \right\rangle \right) \right] = \prod_{j=1}^n \mathbf{E} \left[\exp\left(i \left\langle t, \frac{X_j}{\sqrt{n}} \right\rangle \right) \right] = \left(\Phi_{X_1} \left(\frac{t}{\sqrt{n}} \right) \right)^n.$$

En utilisant le dL du lemme précédent, on obtient une suite (complexe) $\epsilon_n \rightarrow 0$ telle que

$$\Phi_{X_1}\left(\frac{t}{\sqrt{n}}\right) = \left(1 - \frac{\langle t, \mathbf{Cov}(X_1)t \rangle}{2n} + \frac{\epsilon_n}{n}\right) = \left(1 - \frac{\langle t, \mathbf{Cov}(X_1)t \rangle}{2n}\right) \left(1 + \frac{\epsilon_n}{n} \left(1 - \frac{\langle t, \mathbf{Cov}(X_1)t \rangle}{2n}\right)^{-1}\right).$$

Si on pose $\eta_n = \epsilon_n \left(1 - \frac{\langle t, \mathbf{Cov}(X_1)t \rangle}{2n}\right)^{-1} \rightarrow 0$, vérifions que $(1 + \frac{\eta_n}{n})^n \rightarrow 1$ (ce qui est bien connu pour le cas réel en utilisant le log mais peut être moins dans le cas η_n complexe.)= On utilise la formule du binôme :

$$\left| \left(1 + \frac{\eta_n}{n}\right)^n - 1 \right| \leq \sum_{k=1}^n \binom{n}{k} \frac{|\eta_n^k|}{n^k} = \left| \left(1 + \frac{|\eta_n|}{n}\right)^n - 1 \right| \leq |\eta_n| \left(1 + \frac{|\eta_n|}{n}\right)^{n-1} \rightarrow 0,$$

la deuxième inégalité venant du théorème des accroissements finis pour $x > 1$,

$$|f(x) - f(1)| = |(1+x)^n - 1| \leq |x-1| \sup f'(x) = |x-1| n(1+x)^{n-1}.$$

En conclusion

$$\Phi_{\frac{X_1 + \dots + X_n}{\sqrt{n}}}(t) \sim_{n \rightarrow \infty} \left(1 - \frac{\langle t, \mathbf{Cov}(X_1)t \rangle}{2n}\right)^n = \exp\left[n \ln \left(1 - \frac{\langle t, \mathbf{Cov}(X_1)t \rangle}{2n}\right)\right] \rightarrow_{n \rightarrow \infty} \exp\left(-\frac{\langle t, \mathbf{Cov}(X_1)t \rangle}{2}\right).$$

Comme la limite est la fonction caractéristique de la loi gaussienne centrée de covariance $\mathbf{Cov}(X_1)$ le théorème de Paul Lévy conclut. ■

Chapitre 9

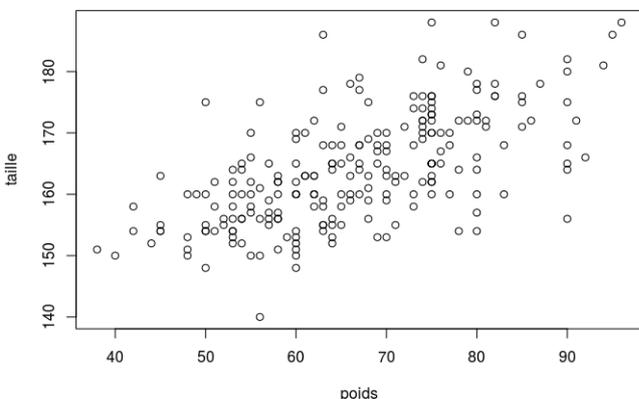
Statistiques descriptives : Deuxième Partie

1 Représentations graphiques bivariées et analyse en composantes principales

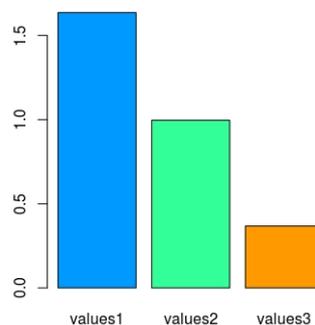
Il est facile de représenter sous R deux variables en même temps sous forme d'un nuage de points. Chaque point représente les coordonnées d'un individu de l'échantillon.

```
plot(taille~poids,main="Nuages de points pour les variables taille et poids")
```

Nuages de points pour les variables taille et poids



Valeurs propres de CSum



Le problème est que pour réaliser cette représentation, il faut avoir seulement deux variables. Plus généralement, on voudrait pouvoir sélectionner un petit nombre de variables plus représentatives. C'est le but de la méthode d'analyse en composantes principales (ACP). Elle permet aussi de trouver des combinaisons linéaires des variables décorréées, et donc plus adaptées pour appliquer un test d'indépendance (bien sûr on peut aussi vouloir connaître l'indépendance des variables de départ).

L'analyse est basée sur la matrice de covariance $(Cov(X_i, X_j))_{i,j}$ ou sur la matrice normalisée pour être sans dimension, la *matrice de Corrélation* (ou matrice de covariances des variables centrées réduites associées) :

$$Cor(X_i, X_j) = \frac{Cov(X_i, X_j)}{\sigma(X_i)\sigma(X_j)}.$$

Remarquez que l'inégalité de Cauchy-Schwartz donne $|Cor(X_i, X_j)| \leq 1$. L'idée est simple, en diagonalisant l'une de ces matrices on obtient des variables décorréées (réduites ou non) et les valeurs

propres donnent une idée des variables les plus aléatoires, qui contiennent l'information principale. Dans la suite on présente la méthode qui diagonalise la matrice de corrélation (ACP de corrélation). Sur des échantillons, on utilise les formules suivantes :

Définition 72. La *covariance empirique non-biaisée* de deux échantillons $x = (x_1, \dots, x_n), y = (y_1, \dots, y_n)$ est la grandeur :

$$cov_n(x, y) = \frac{1}{n-1} \sum_{k=1}^n (x_k - m_n(x))(y_k - m_n(y)).$$

et la *corrélation empirique* est la grandeur :

$$cor_n(x, y) = \frac{cov_n(x, y)}{\sqrt{var_n(x)var_n(y)}}.$$

Pour effectuer l'ACP d'échantillon de variables $x^{(1)} = (x_1^{(1)}, \dots, x_n^{(1)}), \dots, x^{(k)} = (x_1^{(k)}, \dots, x_n^{(k)})$, on considère les données centrées réduites :

$$\hat{x}_j^{(i)} = \frac{x_j^{(i)} - m_n(x^{(i)})}{\sqrt{var_n(x^{(i)})}},$$

la méthode consiste à diagonaliser la matrice (symétrique positive) $C = (cor_n(x^{(i)}, x^{(j)}))_{i,j \in [1,k]} = (cov_n(\hat{x}^{(i)}, \hat{x}^{(j)}))_{i,j \in [1,k]}$.

Les valeurs propres dans l'ordre décroissant permettent de sélectionner les variables nécessaires à représenter 90% ou 75% de la covariance totale. On peut ensuite utiliser la matrice de passage dans la base des vecteurs propres, pour visualiser les données principales.

Dans un petit cas de nos trois variables, on procède ainsi :

```
summary<-data.frame(age,poids,taille);CSum<-cor(summary);EigenSum<-eigen(CSum)
EigenSum[1] renvoie la liste des valeurs propres et M<-matrix(unlist(EigenSum[2]),ncol=3,nrow=3)
```

la matrice dont les colonnes sont les vecteurs propres (matrice de passage de la nouvelle base vers l'ancienne). Si les données sont dans la matrice $A = summary$ 226×3 , on trouve d'abord la matrice $B = SumRed$ des données réduites :

```
Sum2<-matrix(unlist(summary),ncol=3,nrow=226)#Conversion du data.frame en matrice
```

```
SumSD<-apply(Sum2,MARGIN=2,FUN=sd)#Calcul du vecteur d'écart type
```

```
SumM<-colMeans(Sum2)#Calcul du vecteur de moyennes
```

```
SumRed<-t((t(Sum2)-SumM)/SumSD)#Calcul de la matrice de variables réduites
```

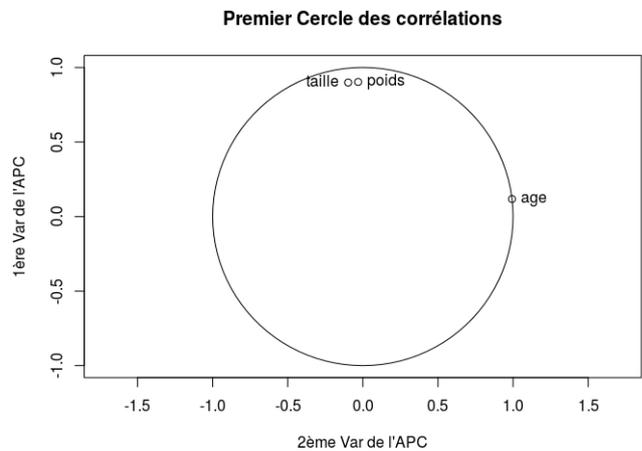
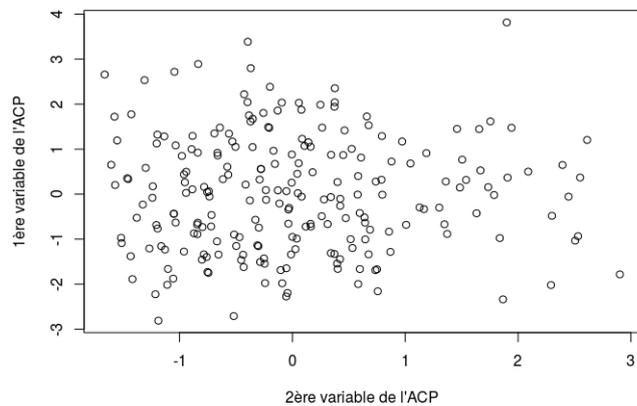
BM est la matrice des données réduites dans la nouvelle base, obtenue par la commande

```
NewSum<-SumRed %*% M %*%diag(c(-1,1,1)); # Ici on change le signe du premier vecteur propre.
```

On peut tracer les deux premières coordonnées (ci-dessous à gauche) par

```
plot(NewSum[,1]~NewSum[,2],ylab="1ère variable de l'ACP",xlab="2ère variable
```

de l'ACP")



On obtient les matrices de corrélations et de passage :

$$C_{Sum} = \begin{pmatrix} 1.00000000 & 0.06248123 & 0.02049272 \\ 0.06248123 & 1.00000000 & 0.63065756 \\ 0.02049272 & 0.63065756 & 1.00000000 \end{pmatrix}, M = \begin{pmatrix} -0.0919494 & 0.99464795 & 0.04712504 \\ -0.7056335 & -0.03169391 & -0.70786780 \\ -0.7025857 & -0.09834102 & 0.70477118 \end{pmatrix}$$

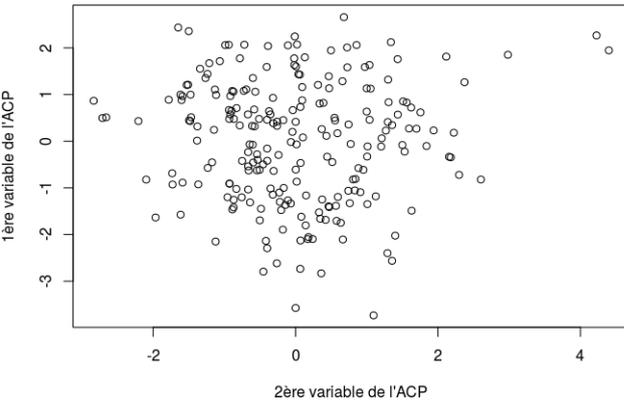
Enfin, on représente souvent les corrélations entre nouvelles variables (x'_1, x'_2, x'_3) et les anciennes variables (age, poids, taille), et on trace les 2 premières coordonnées des 3 vecteurs obtenus pour chaque ancienne coordonnée.

On peut utiliser le code suivant :

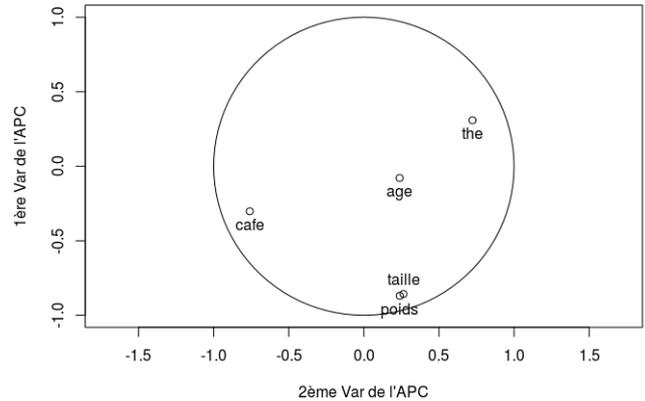
```
library("plotrix")#pour la commande circle
MixCor<-cor(SumRed,NewSum)#Matrice de corrélation entre anciennes et nouvelles variables
plot(MixCor[,1] MixCor[,2],ylim=c(-1,1),xlim=c(-1,1),ylab="1ère Var de l'ACP",xlab="2ème Var de l'ACP",asp = 1,main="Premier Cercle des corrélations")
text(x=MixCor[,2],y=MixCor[,1],colnames(summary),pos=1);draw.circle(0,0,1)
```

On obtient le résultat ci-dessus et ci-dessous si on fait de même pour toutes les variables quantitatives. On voit que dans le premier les 3 coordonnées sont bien capturées par les 2 nouvelles (points près du cercle) et que taille/poids sont très corrélées avec la première nouvelle variable. Dans le deuxième, age est très mal couvert (plus le point est au centre, plus la projection du vecteur de norme 1 est petite, moins bien la variable est représentée), et thé et café sont simultanément pris en compte par la deuxième variable de l'ACP (avec corrélation positive pour thé et négative pour café).

ACP complet de nutriage



Cercle des corrélations complet



2 Introduction à la statistique inférentielle

On présente ici deux thématiques de statistiques plus avancées en continuité avec le cours. On ne vise plus ici à seulement décrire, synthétiser et représenter les observations mais aussi valider ou invalider une théorie/hypothèse, préciser l'erreur autour d'une valeur expérimentale.

2.1 Exemples d'intervalles de confiance pour le paramètre d'une variable de Bernoulli

Pour une suite de v.a. de Bernoulli, la moyenne estime le paramètre, donc la moyenne empirique approche la valeur du paramètre d'un échantillon. Mais étant donné un échantillon, on obtient jamais exactement le paramètre ainsi, il est donc souhaitable d'avoir un intervalle pour lequel on sait avec 90% ou 95% de chance, que le paramètre est contenu dedans.

Ce type d'intervalle est appelé intervalle de confiance de niveau $(1 - \alpha) = 0.9$ ou $(1 - \alpha) = 0.95$. Donnons 3 exemples de moyens d'obtenir ce résultat.

Exemple 61. Intervalle de confiance exact. L'inégalité de Tchebychev donne un tel intervalle de confiance, en effet, on sait que pour $S_n = X_1 + \dots + X_n$ somme de v.a. i.i.d $B(1, p)$, on a $E(S_n) = np$, $Var(S_n) = np(1 - p)$ donc l'inégalité donne :

$$P\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) = P(|S_n - np| \geq n\epsilon) \leq \frac{Var(S_n)}{n^2\epsilon^2} \leq \frac{1}{4n\epsilon^2}$$

Pour obtenir une valeur inférieure à α , on prend $\epsilon_n = \frac{1}{\sqrt{4n\alpha}}$ de sorte que $\frac{1}{4n\epsilon^2} = \alpha$. Donc avec proba supérieur à α

$$p \in \left[\frac{S_n}{n} - \frac{1}{\sqrt{4n\alpha}}, \frac{S_n}{n} + \frac{1}{\sqrt{4n\alpha}}\right]$$

c'est l'intervalle de confiance de niveau α cherché.

Exemple 62. Premier Intervalle de confiance asymptotique. On obtient des intervalles plus étroits si on utilise des théorèmes limites valables seulement pour n grand. Le Thm de Moivre Laplace donne

$$P\left(\frac{S_n}{n} \in [p - \sigma T/\sqrt{n}, p + \sigma T/\sqrt{n}]\right) = P\left(\frac{S_n - np}{\sqrt{n}\sigma} \in [-T, T]\right) \rightarrow \int_{-T}^T e^{-x^2/2} \frac{1}{\sqrt{2\pi}} dx$$

Soit $z_{\alpha/2}$ le fractile d'ordre $1 - \alpha/2$ de la loi normale $\mathcal{N}(0, 1)$ définie (pour $Y \sim \mathcal{N}(0, 1)$) par $P(Y \leq z_{\alpha/2}) = 1 - \alpha/2$ de sorte que $P(Y \geq z_{\alpha/2}) = \alpha/2$, $P(|Y| \geq z_{\alpha/2}) = \alpha$ par symétrie et donc

$$P\left(\frac{S_n}{n} \in [p - \sigma z_{\alpha/2}/\sqrt{n}, p + \sigma z_{\alpha/2}/\sqrt{n}]\right) = P\left(\frac{S_n - np}{\sqrt{n}\sigma} \in [-T, T]\right) \rightarrow 1 - \alpha$$

On sait donc qu'avec probabilité de l'ordre de $1 - \alpha$, pour n grand, on a $p \in [\frac{S_n}{n} - \sigma z_{\alpha/2}/\sqrt{n}, \frac{S_n}{n} + \sigma z_{\alpha/2}/\sqrt{n}]$ mais cela ne donne pas encore un intervalle de confiance vu que $\sigma^2 = p(1 - p)$ dépend de p . Mais comme avant $\sigma \leq 1/4$ d'où l'intervalle de confiance d'ordre $1 - \alpha$:

$$p \in \left[\frac{S_n}{n} - \frac{z_{\alpha/2}}{\sqrt{4n}}, \frac{S_n}{n} + \frac{z_{\alpha/2}}{\sqrt{4n}} \right].$$

On peut retenir $z_{0.025} \simeq 1.96 < \frac{1}{\sqrt{0.05}} \simeq 4.47$, $z_{0.05} \simeq 1,65 < \frac{1}{\sqrt{0.1}} \simeq 3.16$.

Exemple 63. Deuxième Intervalle de confiance asymptotique. Un raffinement du TCL, le lemme de Slutsky dit que si σ_n^2 est un estimateur de la variance convergeant avec proba 1 (comme donné par la LGN) alors la conclusion du TCL s'applique à $\frac{S_n - np}{\sqrt{n}\sigma_n}$. Ainsi dans le cas bernoulli, la LGN donne $\sigma_n^2 = \frac{S_n}{n}(1 - \frac{S_n}{n})$ et l'intervalle de confiance d'ordre $1 - \alpha$ (meilleur pour p loin de $1/2$) :

$$p \in \left[\frac{S_n}{n} - \frac{z_{\alpha/2} \sqrt{\frac{S_n}{n}(1 - \frac{S_n}{n})}}{\sqrt{n}}, \frac{S_n}{n} + \frac{z_{\alpha/2} \sqrt{\frac{S_n}{n}(1 - \frac{S_n}{n})}}{\sqrt{n}} \right].$$

2.2 Un exemple de test d'hypothèse : le test d'indépendance du χ^2

Problématique

On a vu plus haut comment se ramener à des variables presque décorréées par l'ACP, mais cela ne suffit pas à savoir que les variables sont presque indépendantes ou non. De même que pour les paramètres de variables de Bernoulli, des échantillons statistiques ne donneront jamais des tableaux de contingences de variables exactement indépendantes comme vu en TD. On introduit ici une thématique générale de statistique, celle des tests d'hypothèse, qui vise à quantifier l'erreur sur une hypothèse, ici celle d'indépendance et dire que par exemple avec 95% de chances de ne pas se tromper, les variables ne sont pas indépendantes mais si on autorise seulement 90% de chance de ne pas se tromper, elles le sont presque.

Commençons par souligner un problème général. On a une hypothèse à tester H_0 , par exemple, les variables sexe et matgras de notre jeu de données sont indépendantes et son hypothèse contraire H_1 , elles ne sont pas indépendantes. Si on pense qu'elles sont indépendantes si elles le sont ou qu'elles ne le sont pas si elles ne le sont pas, il n'y a pas de problème. Mais on a deux possibilités d'erreurs, soit on croit qu'elles sont indépendantes alors qu'elles ne le sont pas (faux positif ou erreur de seconde espèce : on accepte H_0 alors que H_0 est faux) soit on pense qu'elles ne sont pas indépendantes alors qu'elles le sont (faux négatif ou erreur de première espèce : on rejette H_0 alors que H_0 est vrai).

Les deux risques d'erreurs de première espèce α et de seconde espèce β sont en général différents, et un test de H_0 contre H_1 quantifie surtout le premier risque α . (Si il existe aussi un test de H_1 contre H_0 , ce qui n'est pas toujours le cas, les types de risques s'inversent, et on peut alors quantifier plus l'autre risque). Techniquement, on a différentes lois de proba P_x paramétrant les possibles $H_0 \subset X$. On a donc $\alpha = \sup_{x \in H_0} P_x(\text{rejeter } H_0)$ est le seuil du test et quand on peut le calculer $\beta = \sup_{x \in H_1} P(\text{accepter } H_0)$ donne la puissance du test $(1 - \beta)$. On prend souvent 5% pour α . **Un test de H_0 permet de rejeter H_0 avec bonne précision, quand on ne rejette pas H_0 , on ne sait pas que H_0 est vrai, juste qu'on a un risque d'erreur β qui ne sera pas quantifiée dans ce cours.** Ce qu'on a minimisé, c'est l'erreur α d'avoir rejeté H_0 à tort. Ce choix est normal car le but de toute science expérimentale est de falsifier ces hypothèses fausses (cf Karl Popper *La logique de la découverte scientifique*). De plus, les tests ne disent rien sur la pertinence de la question posée...

Test d'indépendance de variables discrètes (nominales ou quantitatives discrètes)

Si on a les variables X, Y à valeurs I, J finies respectivement et qu'on a les effectifs empiriques $n = (n_{i,j})$ ($n_{i,j}$ nombre d'individus de l'échantillon avec $X = i$ et $Y = j$) et l'effectif total $N = \sum_{i \in I} \sum_{j \in J} n_{i,j}$. On peut calculer les effectifs marginaux $N_{(i,\cdot)} = \sum_{j \in J} n_{i,j}$ et $N_{(\cdot,j)} = \sum_{i \in I} n_{i,j}$. Si les caractères étaient indépendants, on aurait $\frac{n_{i,j}}{N} = \frac{N_{(i,\cdot)}N_{(\cdot,j)}}{N^2}$. Ce qui donne les effectifs théoriques sous l'hypothèse vraie : $N_{i,j}^{th} = \frac{N_{(i,\cdot)}N_{(\cdot,j)}}{N}$. Le test va donc quantifier l'écart à la validité de cette équation. On considère la statistique de Pearson :

$$D^2(n) = \sum_{i \in I} \sum_{j \in J} \frac{(n_{i,j} - N_{i,j}^{th})^2}{N_{i,j}^{th}}$$

Un théorème limite du niveau L3 dit que cette grandeur tend (en loi c'est à dire dans le même sens que dans le TCL) vers une loi $\chi^2(d)$ avec $d = (Card(I) - 1)(Card(J) - 1)$ est appelé nombre de degrés de libertés. Le test dit donc qu'au niveau α si $\chi_\alpha^2(d)$ est le quantile tel que $P(\chi^2(d) \leq \chi_\alpha^2(d)) = 1 - \alpha$ et :

1. si $D^2(n) > \chi_\alpha^2(d)$ on rejette l'hypothèse d'indépendance.
2. si $D^2(n) \leq \chi_\alpha^2(d)$ on ne peut pas rejeter l'hypothèse d'indépendance au niveau α , et **par ABUS** on dit parfois qu'on l'accepte.

En pratique, comme le test est asymptotique, il faut au moins que $N \geq 30$ et tous les $N_{i,j} \geq 5$ pour pouvoir appliquer le test (sinon on regroupe les classes).

Exemple 64. Appliquons au cas de l'exemple de la section 1.1 dont la table de contingence *TCsexeMatgras* a déjà été représentée :

	matgras							
sexe	beurre	margarine	arachide	tournesol	olive	Melange	colza	canard
Femme	5	17	32	47	20	18	1	1
Homme	10	10	16	21	20	5	0	3

La commande suivante donne les effectifs théoriques

`ChiIndepSexeMatgra <- chisq.test(TCsexeMatgras); ChiIndepSexeMatgra$expected :`

	matgras							
sexe	beurre	margarine	arachide	tournesol	olive	Melange	colza	canard
Femme	9.358407	16.84513	29.9469	42.42478	24.95575	14.349558	0.6238938	2.495575
Homme	5.641593	10.15487	18.0531	25.57522	15.04425	8.650442	0.3761062	1.504425

Les dernières classes ont moins de cinq éléments, ce qui nécessite de rassembler des classes. On le réalise ainsi

```
matgrasReduit<-matgras
levels(matgrasReduit)<-c("beurreOucanard","margarine","arachideOUcolza","tournesol","olive",
,"arachideOUcolza", "beurreOucanard")
TableClassesChiSQ<-table(sexe,matgrasReduit)
ChiIndepSexeMatgra2<-chisq.test(TableClassesChiSQ) ; ChiIndepSexeMatgra2$expected
ce qui donne :
```

	<i>matgras</i>					
<i>sexe</i>	<i>beurreOuCanard</i>	<i>margarine</i>	<i>arachideOucolza</i>	<i>tournesol</i>	<i>olive</i>	<i>Melange</i>
<i>Femme</i>	11.853982	16.84513	30.5708	42.42478	24.95575	14.349558
<i>Homme</i>	7.146018	10.15487	18.4292	25.57522	15.04425	8.650442

Le calcul de la statistique de Pearson donne $\text{sum}((\text{ChiIndepSexeMatgra2}\$residuals)^2)$ [1] **14.60111** puis le quantile pour $(6-1)(2-1)=5$ degrés de libertés : $\text{qchisq}(0.95,df=5)$ [1] **11.0705** comme $14.60111 > 11.0705$ on rejette l'hypothèse d'indépendance au niveau $\alpha = 5\%$. Remarquez que $\text{qchisq}(0.99,df = 5) = 15.08627 > 14.60111$ donc on ne rejeterait pas l'hypothèse d'indépendance si on n'autorisait qu'une erreur de 1% sur la proba de penser qu'ils ne sont pas indépendants si ils le sont. Plus le α est petit, plus il est difficile de rejeter H_0 , et moins de fois le test est concluant puisque le test est informatif quand on rejette l'hypothèse. Plus généralement en calculant $1 - \text{pchisq}(14.60111,df = 5) = 0.01220989$ (le risque de première espèce maximal permettant de ne pas rejeter l'hypothèse) et on rejette l'hypothèse d'indépendance puisque la valeur est inférieure à 5%. Plus ce nombre est petit, plus il est probant de rejeter l'hypothèse.

Exemple 65. On a vu à l'analyse en composante principale que taille et age semblaient peu corrélés. On peut appliquer le test d'indépendance pour confirmer cela une fois qu'on a réparti les effectifs par classes.

Vous pouvez vous entraîner en programmant une fonction pour obtenir des variables *NewAge*, *NewTaille* donnant les données par classes. On applique alors le code suivant :

```
NewTailleAge<-data.frame(NewAge,NewTaille)
TableAgeTaille<-table(NewTailleAge) ;TableAgeTaille
```

	<i>NewTaille</i>				
<i>NewAge</i>	140 – 155	155 – 160	160 – 165	165 – 170	170 – 190
65 – 70	11	8	17	8	8
70 – 75	11	11	17	8	26
75 – 80	8	11	13	12	20
80 – 95	6	7	7	7	10

```
ChiSQTailleAge<-chisq.test(TableAgeTaille)
on obtient les effectifs attendues : ChiSQTailleAge$expected
```

	<i>NewTaille</i>				
<i>NewAge</i>	140 – 155	155 – 160	160 – 165	165 – 170	170 – 190
65 – 70	8.283186	8.513274	12.424779	8.053097	14.72566
70 – 75	11.628319	11.951327	17.442478	11.305310	20.67257
75 – 80	10.194690	10.477876	15.292035	9.911504	18.12389
80 – 95	5.893805	6.057522	8.840708	5.730088	10.47788

Les classes ont toutes au moins un effectif de 5, donc on peut appliquer le test, on obtient la statistique de Pearson et la p-valeur :

`sum((ChiSQTailleAge$residuals)2)[1]` 10.45048

On a $(5 - 1) * (4 - 1) = 12$ degrés de liberté donc la p-valeur est : $1 - pchisq(10.45048, df = 12) = 0.5765037 > 0,1$ donc il n'y a pas de présomption contre l'hypothèse d'indépendance. On ne rejette pas l'hypothèse. Cela veut dire qu'on aurait plus d'une chance sur 2 de se tromper en rejetant l'hypothèse si elle était vraie.

Chapitre 10

Théorème de Moivre-Laplace II : Démonstration

1 Formule de Stirling

On commence par montrer la formule de Stirling dont on va avoir besoin dans la section suivante.

Lemme 75. (*Formule de Stirling*)

$$n! \sim \sqrt{2\pi n} e^{-n} n^n.$$

Plus précisément, on a l'encadrement :

$$\sqrt{2\pi n} e^{-n} n^n < n! < \sqrt{2\pi n} e^{-n + \frac{1}{12n}} n^n.$$

Démonstration. On compare les log des 2 termes :

$$d_n = \log(n!) - \log(\sqrt{n} e^{-n} n^n) = \log(n!) - \left(n + \frac{1}{2}\right) \log(n) + n.$$

On va écrire ceci comme une série télescopique en calculant

$$d_n - d_{n+1} = \left(n + \frac{1}{2}\right) \log\left(\frac{n+1}{n}\right) + \log(n+1) - \log(n+1) - 1 = \frac{1}{2t_n} \log\left(\frac{1+t_n}{1-t_n}\right) - 1 =: f(t_n)$$

avec $t_n = \frac{1}{2n+1}$. En fait un DL donne que c'est un $O(1/n^2)$ ce qui suffit à voir que la série converge par comparaison aux séries de Riemann, d'où la suite d_n converge par critère des séries télescopiques. On va obtenir un encadrement plus précis en utilisant l'inégalité sur f que l'on pourrait obtenir par le développement en série entière en la variable $t_n \leq 1$. On montre l'inégalité directement, pour $t \in [0, 1]$,

$$f(t) = \frac{1}{2t} \log\left(\frac{1+t}{1-t}\right) - 1 \leq \frac{t^2}{3(1-t^2)} = g(t).$$

En effet, pour $0 < t < 1$, on a le développement en série entière (intégration de la série géométrique) :

$$\begin{aligned} \ln(1-x) &= \sum_{n=1}^{\infty} \frac{x^n}{n} \\ \ln(1+x) &= \sum_{n=1}^{\infty} \frac{(-1)^n x^n}{n} \end{aligned}$$

soit en soustrayant :

$$\log\left(\frac{1+t}{1-t}\right) = 2 \sum_{n=0}^{\infty} \frac{t^{2n+1}}{2n+1}$$

$$0 < f(t) = \sum_{n=1}^{\infty} \frac{t^{2n}}{2n+1} < \frac{1}{3} \sum_{n=1}^{\infty} t^{2n} = g(t).$$

En évaluant en t_n on obtient : $0 < d_n - d_{n+1} < \frac{1}{3[(2n+1)^2-1]} = \frac{1}{3[(2n+2)2n]} = \frac{1}{12n} - \frac{1}{12(n+1)}$

Donc (d_n) est décroissante, $d_n - \frac{1}{12n}$ est croissante et la différence tend vers 0, les suites sont donc adjacentes, et ont donc même limite, disons C , telle que $d_n - \frac{1}{12n} < C < d_n$. En prenant l'exponentielle cela se traduit en :

$$\sqrt{n}e^{C-n}n^n < n! < \sqrt{n}e^{C-n+\frac{1}{12n}}n^n.$$

Il reste donc à voir $e^C = \sqrt{2\pi}$. On pourra voir en suivant les constantes dans le Théorème de Moivre Laplace que la constante doit être la normalisation de la gaussienne. L'autre méthode classique est d'utiliser les intégrales de Wallis (cf e.g. Gourdon d'analyse p 127) donnant la limite classique :

$$\lim_{p \rightarrow \infty} \frac{1}{p} \frac{4^{2p}(p!)^4}{((2p)!)^2} = \pi.$$

En remplaçant dans l'équivalent précédent, on trouve la constante voulue. □

2 Théorème de Moivre-Laplace proprement dit

On va montrer le résultat suivant dont on verra des applications en statistique au chapitre suivant :

Théorème 76. [Théorème de Moivre-Laplace] Soit $0 < p < 1$. Soit $(X_i)_{i \geq 1}$ une suite de v.a. de Bernoulli $B(1, p)$ indépendantes et identiquement distribuées (donc de moyenne $m = p$ et de variance $\sigma^2 = p(1-p)$) et soit Y une v.a. de loi normale $\mathcal{N}(0, \sigma^2)$, alors

$$P\left(\left\{\omega : \frac{X_1(\omega) + \dots + X_n(\omega) - nm}{\sqrt{n}} \in [a, b]\right\}\right) \xrightarrow{n \rightarrow \infty} P(Y \in [a, b]) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2}} dt.$$

Concrètement, le membre de gauche se calcule ainsi comme pour toute loi discrète, vu $S_n = X_1 + \dots + X_n$ de loi binomiale :

$$P_n([a, b]) := P\left(\left\{\omega : \frac{X_1(\omega) + \dots + X_n(\omega) - nm}{\sqrt{n}} \in [a, b]\right\}\right) = \sum_{k \in [a, b]} P(S_n = np + k\sqrt{n})$$

où la somme ne contient que les k tel que $np + k\sqrt{n}$ est entier pour que la probabilité indiquée soit non nulle.

Avant de montrer le théorème, on va donc trouver un équivalent de chacune des probabilités dans la somme (en utilisant la formule de Stirling). On rassemblera ensuite ces termes pour obtenir une somme de Riemann. On garde les notations précédentes :

Lemme 77. (de limite locale) On a la borne donnant un équivalent uniforme (autour de la moyenne) :

$$\sup_{|x| \leq \psi(n)} \left| \frac{P\left(\frac{S_n - np}{\sqrt{n}} = x\right)}{\frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{x^2}{2p(1-p)}}} - 1 \right| \xrightarrow{n \rightarrow \infty} 0$$

pour tout $\psi(n)$ tel que $\psi(n)/\sqrt{n} \rightarrow 0$.

Démonstration. L'équivalent de la formule de Stirling donne une suite $\epsilon_n \in [0, \epsilon'_n]$ $\epsilon'_n = \exp(1/12n) - 1 \rightarrow 0$ tel que $n! = n^n e^{-n} \sqrt{2\pi n} (1 + \epsilon_n)$. On a aussi $\psi(n) = \eta_n \sqrt[6]{n}$, $\eta_n \rightarrow 0$ Soit $k = x\sqrt{n}$ et rappelons que pour la loi binomiale : $P(\frac{S_n - np}{\sqrt{n}} = x) = \binom{n}{np+k} p^{np+k} (1-p)^{n(1-p)-k}$.

On remplace donc la formule de Stirling dans le coefficient binomial ($q = 1 - p$) :

$$\begin{aligned} \binom{n}{np+k} &= \frac{n^n e^{-n} \sqrt{2\pi n} (1 + \epsilon_n)}{(np+k)^{(np+k)} e^{-(np+k)} \sqrt{2\pi(np+k)} (1 + \epsilon_{(np+k)}) (nq-k)^{(nq-k)} e^{-(nq-k)} \sqrt{2\pi(nq-k)} (1 + \epsilon_{(nq-k)})} \\ &= \frac{(1 + \epsilon_n)}{(p + \frac{k}{n})^{(np+k)} \sqrt{2\pi n} (p + \frac{k}{n}) (1 + \epsilon_{(np+k)}) (q - \frac{k}{n})^{(nq-k)} \sqrt{q - \frac{k}{n}} (1 + \epsilon_{(nq-k)})} \end{aligned}$$

de sorte que :

$$\binom{n}{np+k} p^{np+k} q^{nq-k} \sqrt{2\pi np(1-p)} = \frac{(1 + \epsilon_n)}{(1 + \epsilon_{(np+k)}) (1 + \epsilon_{(nq-k)})} \frac{1}{(1 + \frac{k}{pn})^{(np+k)} (1 - \frac{k}{qn})^{(nq-k)} \sqrt{(1 + \frac{k}{pn})(1 - \frac{k}{qn})}}$$

D'abord on a les inégalités

$$\frac{1}{\exp\left(\frac{1}{12} \left(\frac{n}{(np+k)(nq-k)}\right)\right)} = \frac{1}{\exp\left(\frac{1}{12} \left(\frac{1}{np+k} + \frac{1}{nq-k}\right)\right)} \leq \frac{(1 + \epsilon_n)}{(1 + \epsilon_{(np+k)}) (1 + \epsilon_{(nq-k)})} \leq \exp\left(\frac{1}{12n}\right) \quad (10.1)$$

qui donne des bornes convergeant uniformément en $|x| = |k/\sqrt{n}| \leq \psi(n)$ vers 1. Idem les termes en $\sqrt{(1 + \frac{k}{pn})(1 - \frac{k}{qn})}$ vu $|k/n| \leq \psi(n)/\sqrt{n} = o(n^{-1/3})$.

Or $(1 + \frac{k}{pn})^{(np+k)} (1 - \frac{k}{qn})^{(nq-k)} = \exp((np+k) \ln(1 + \frac{k}{pn}) + (nq-k) \ln(1 - \frac{k}{qn})) = \exp(-nH(p + \frac{k}{n}))$ avec $H(y) = y \ln(\frac{y}{p}) + (1-y) \ln(\frac{1-y}{q})$.

Pour appliquer la formule de Taylor à l'ordre 3, on remarque que H est C^3 pour $y \in]0, 1[$,

$$H'(y) = \ln\left(\frac{y}{p}\right) - \ln\left(\frac{1-y}{q}\right) + y/y - (1-y)/(1-y)$$

$$H''(y) = \frac{1}{y} + \frac{1}{1-y} = \frac{1}{y(1-y)}$$

$$H'''(y) = -\frac{1}{y^2} + \frac{1}{(1-y)^2}$$

On applique donc Taylor avec reste intégral en $y = p$ (notez $H(p) = H'(p) = 0$)

$$H(p + \frac{k}{n}) = H(p) + H'(p) \frac{k}{n} + H''(p) \frac{k^2}{2n^2} + \int_p^{p+\frac{k}{n}} H'''(x) \frac{(x-p)^2}{2} dx$$

donc pour n assez grand tel que $[p, p + k/n] \subset [p - \epsilon, p + \epsilon]$ fixé de sorte que H''' borné par C sur cet intervalle, on a :

$$\left| nH(p + \frac{k}{n}) - \frac{x^2}{2pq} \right| \leq Cn \frac{k^3}{6n^3} \leq C\eta_n^3$$

uniformément sur $|x| = |k/\sqrt{n}| \leq \psi(n) = \eta_n n^{1/6}$.

En rassemblant toutes les bornes uniformes, on obtient le résultat. \square

Démonstration du Théorème 76. On part de la formule :

$$P_n([a, b]) = \sum_{k:t_k \in [a, b]} P(S_n = np + t_k \sqrt{n})$$

où on a posé t_k tel que $k = np + t_k \sqrt{n}$ avec k entier. Noter que $\Delta t_k = t_k - t_{k-1} = 1/\sqrt{n}$. Le lemme précédent montre que

$$P(S_n = np + t_k \sqrt{n}) = \frac{1}{\sqrt{2\pi p(1-p)}} e^{-\frac{t_k^2}{2p(1-p)}} \Delta t_k (1 + \epsilon(t_k, n))$$

avec $\sup_{|t_k| \leq T} |\epsilon(t_k, n)| \rightarrow_{n \rightarrow \infty} 0$ pour T fixé avec $[a, b] \subset [-T, T]$.

$$P_n([a, b]) = \sum_{k:t_k \in [a, b]} \frac{1}{\sqrt{2\pi p(1-p)}} e^{-\frac{t_k^2}{2p(1-p)}} \Delta t_k + \sum_{k:t_k \in [a, b]} \frac{1}{\sqrt{2\pi p(1-p)}} e^{-\frac{t_k^2}{2p(1-p)}} \Delta t_k \epsilon(t_k, n)$$

La première somme est une somme de Riemann qui converge vers l'intégrale voulue. Il reste à montrer que la deuxième somme tend vers 0. Or le deuxième terme est borné par

$$\sup_{|t_k| \leq T} |\epsilon(t_k, n)| \sum_{k:t_k \in [a, b]} \frac{1}{\sqrt{2\pi p(1-p)}} e^{-\frac{t_k^2}{2p(1-p)}} \Delta t_k \rightarrow_{n \rightarrow \infty} 0$$

car la deuxième suite est la somme de Riemann précédente qui est bornée car convergente.

Mais en fait comme on n'a pas encore justifié la constante dans la formule de Stirling, on n'a obtenu le résultat que pour $\sqrt{2\pi}$ remplacé par e^C dans le théorème.

Pour calculer la constante et finir cette preuve et celle de la formule de Stirling. On calcule la limite de

$$1 = P\left(\left\{\omega : \frac{X_1(\omega) + \dots + X_n(\omega) - nm}{\sqrt{n}} \in \mathbb{R}\right\}\right) \rightarrow_{n \rightarrow \infty} \int_{\mathbb{R}} \frac{1}{\sigma e^C} e^{-\frac{t^2}{2\sigma^2}} dt.$$

Ceci force $e^C = \sqrt{2\pi}$ par le calcul de la constante dans la densité de la loi gaussienne.

Pour appliquer le théorème de limite locale, on décompose avec $\psi(n) = n^{1/12}$:

$$\begin{aligned} P\left(\left\{\omega : \frac{X_1(\omega) + \dots + X_n(\omega) - nm}{\sqrt{n}} \in \mathbb{R}\right\}\right) &= P\left(\left\{\omega : |X_1(\omega) + \dots + X_n(\omega) - nm| \geq \sqrt{nn}^{1/12}\right\}\right) \\ &+ \sum_{k:t_k \in [-\psi(n), \psi(n)]} \frac{1}{\sqrt{2\pi p(1-p)}} e^{-\frac{t_k^2}{2p(1-p)}} \Delta t_k + \sum_{k:t_k \in [-\psi(n), \psi(n)]} \frac{1}{\sqrt{2\pi p(1-p)}} e^{-\frac{t_k^2}{2p(1-p)}} \Delta t_k \epsilon(t_k, n) \end{aligned}$$

Le premier terme se borne par Tchebychev :

$$P\left(\left\{\omega : |X_1(\omega) + \dots + X_n(\omega) - nm| \geq \sqrt{nn}^{1/12}\right\}\right) \leq \frac{\text{Var}(X_1 + \dots + X_n)}{nn^{1/6}} = \frac{\text{Var}(X_1)}{n^{1/6}} \rightarrow_{n \rightarrow \infty} 0$$

Le deuxième terme se compare à

$$\int_{-\psi(n)}^{\psi(n)} \frac{1}{\sigma e^C} e^{-\frac{t^2}{2\sigma^2}} dt \rightarrow \int_{-\infty}^{\infty} \frac{1}{\sigma e^C} e^{-\frac{t^2}{2\sigma^2}} dt$$

et comme auparavant par intégrale de Riemann, on a (avec $K = \lceil pn - \sqrt{n}\psi(n) \rceil$ de sorte que $k \geq K$ dans la somme ci-dessous)

$$\begin{aligned}
& \left| \sum_{k:t_k \in [-\psi(n), \psi(n)]} \frac{1}{\sqrt{2\pi p(1-p)}} e^{-\frac{t_k^2}{2p(1-p)}} \Delta t_k - \int_{-\psi(n)}^{\psi(n)} \frac{1}{\sigma e^C} e^{-\frac{t^2}{2\sigma^2}} dt \right| \\
& \leq \sum_{k:t_k \in [-\psi(n), \psi(n)]} \frac{1}{\sqrt{2\pi p(1-p)}} \left(\int_{-\psi(n)+(k-K)/\sqrt{n}}^{-\psi(n)+(k-K+1)/\sqrt{n}} (e^{-\frac{t_k^2}{2p(1-p)}} - e^{-\frac{s^2}{2p(1-p)}}) ds \right) \\
& \leq (2\psi(n)\sqrt{n} + 3) \frac{1}{\sqrt{2\pi p(1-p)}} \frac{1}{\sqrt{n}} \frac{2}{\sqrt{n}} (\psi(n) + 1/\sqrt{n}) \frac{1}{p(1-p)} \leq \frac{C\psi(n)^2}{\sqrt{n}} \rightarrow 0
\end{aligned}$$

vu la borne par le Théorème des accroissements finis

$$\left| e^{-\frac{t_k^2}{2p(1-p)}} - e^{-\frac{s^2}{2p(1-p)}} \right| \leq |t_k - s| \sup_{u \in [t_k, s]} \frac{ue^{-\frac{u^2}{2p(1-p)}}}{p(1-p)} \leq \frac{2}{\sqrt{n}} (|t_k| + 1/\sqrt{n}) \frac{1}{p(1-p)}$$

(vu $|s| \leq |t_k| + 1/\sqrt{n}$)

Le dernier terme a une borne uniforme par le Théorème de limite locale vu $\psi(n)/n^{1/6} \rightarrow 0$. \square