

Licence 2ème année.

PROBABILITÉ ET STATISTIQUES.

par **Bruno SCHAPIRA**, et légèrement modifiée par **Maxime Hauray**

(d'après les notes de cours d'Amine Asselah et Fabienne Castell.)

Table des matières

| | | |
|----------|--|-----------|
| 1 | Issues, Probabilités, Variables aléatoires. | 1 |
| 1.1 | Introduction | 1 |
| 1.1.1 | Hasard et information incomplète? | 1 |
| 1.2 | Espace des issues. | 2 |
| 1.3 | Probabilités. | 4 |
| 1.4 | Variables aléatoires | 6 |
| 1.4.1 | Définition | 6 |
| 1.4.2 | Loi d'une variable aléatoire | 6 |
| 1.4.3 | Fonction de répartition | 8 |
| 1.4.4 | Fonction indicatrice. | 9 |
| 2 | Espérance, variance, écart type. | 11 |
| 2.1 | Espérance. | 11 |
| 2.1.1 | Définitions. | 11 |
| 2.1.2 | Propriétés de l'espérance. | 13 |
| 2.2 | Variance, covariance, et écart type. | 13 |
| 2.3 | Espérance et variance de quelques lois classiques. | 15 |
| 3 | Indépendance, Probabilités Conditionnelles. | 19 |
| 3.1 | Indépendance. | 19 |
| 3.2 | Probabilités Conditionnelles. | 23 |
| 3.2.1 | Motivation. | 23 |
| 3.2.2 | Probabilité conditionnelle. | 23 |
| 3.2.3 | Formules des probabilités totales et des probabilités composées. | 25 |
| 4 | Fonction génératrice. | 27 |
| 4.1 | Fonction génératrice. | 27 |
| 4.2 | Somme d'un nombre aléatoire de variables aléatoires. | 29 |
| 4.3 | Evolution des Populations | 31 |
| 4.3.1 | Motivation | 31 |
| 4.3.2 | Probabilité d'extinction. | 31 |
| 4.3.3 | Exemple. | 32 |
| 5 | Variables continues. Loi normale. | 33 |
| 5.1 | Axiomatique des probabilités. | 33 |
| 5.2 | Variables aléatoires. | 34 |
| 5.2.1 | Définition. | 34 |

| | | |
|-----------|---|-----------|
| 5.2.2 | Variables à densité. | 35 |
| 5.2.3 | Indépendance. | 37 |
| 5.2.4 | Espérance. | 37 |
| 5.3 | Lois Normales | 38 |
| 6 | Théorèmes limites. | 41 |
| 6.1 | Loi des grands nombres. | 41 |
| 6.1.1 | Motivation. | 41 |
| 6.1.2 | Espérance et variance d'une moyenne empirique. | 42 |
| 6.1.3 | Inégalités de Markov et de Chebyshev. | 43 |
| 6.1.4 | Loi faible des grands nombres. | 43 |
| 6.2 | Convergence en loi | 44 |
| 6.3 | Théorème central limite. | 45 |
| 6.3.1 | Énoncé général | 45 |
| 6.3.2 | Approximations de la loi binomiale | 46 |
| 7 | Concepts de base de la statistique. | 47 |
| 7.1 | Modèles Statistiques | 47 |
| 7.2 | Quel genre d'informations tirer d'un modèle statistique? | 48 |
| 7.3 | Estimation | 48 |
| 7.3.1 | Estimateur. | 48 |
| 7.3.2 | Construction d'estimateurs | 49 |
| 7.3.3 | Biais et risque d'un estimateur | 51 |
| 8 | Intervalles de confiance | 53 |
| 8.1 | Echantillon de Bernoulli | 53 |
| 8.2 | Définitions générales | 54 |
| 8.3 | Échantillons gaussiens | 55 |
| 8.3.1 | Estimateurs de la moyenne et de la variance. | 55 |
| 8.3.2 | Intervalle de confiance pour μ avec σ connu. | 56 |
| 8.3.3 | Intervalle de confiance pour μ avec σ inconnu. | 56 |
| 8.3.4 | Intervalle de confiance pour σ^2 | 57 |
| 9 | Premiers tests : échantillons de Bernoulli et échantillons gaussiens | 59 |
| 9.1 | Échantillons de Bernoulli | 59 |
| 9.2 | Échantillons gaussiens. | 63 |
| 9.2.1 | Test sur μ avec σ connu. | 63 |
| 9.2.2 | Test sur μ avec σ inconnu. | 64 |
| 9.2.3 | Test sur σ | 65 |
| 10 | Tests du χ^2 | 67 |
| 10.1 | Test du χ^2 d'ajustement. | 67 |
| 10.1.1 | Test du χ^2 d'ajustement à une loi sur un espace fini. | 67 |
| 10.1.2 | Test du χ^2 d'ajustement à une loi quelconque | 68 |
| 10.1.3 | Test du χ^2 d'ajustement à une famille de loi. | 69 |
| 10.2 | Test du χ^2 d'indépendance. | 71 |

Chapitre 1

Issues, Probabilités, Variables aléatoires.

1.1 Introduction

1.1.1 Hasard et information incomplète ?

Un exemple classique : un jeu de dé. On secoue un dé dans un gobelet, fermé par la main, et on le lance ensuite sur une table. Comme d’habitude on regarde la face qui se retrouve en haut. Ceci est un problème de mécanique. Connaissant la façon dont il a été lancé et secoué, on pourrait penser prévoir le résultat du lancer du dé. Mais c’est impossible car :

- il est très difficile de savoir exactement comment un dé a été lancé et secoué,
- les rebonds d’un dés sont très dur à prévoir : une très petite modification de l’angle formé avec la paroi au moment du rebond peut grandement modifier la trajectoire après le choc.

Et donc il est difficile de prévoir le mouvement du dés après quelques rebonds, c’est pour cela qu’il est important de bien le secouer (ou de le lancer fort).

Par contre, si le dé n’est pas pipé (c’est-à-dire s’il est bien symétrique), et qu’on le lance avec suffisamment de rebond, on peut penser que toutes les faces vont toutes “sortir” en moyenne avec la même fréquence : $1/6$.

Ici, on ne peut pas décrire précisément ce qui se passe, il nous manque des informations : Il y a du **hasard**. Mais on connaît les fréquences moyennes de chaque face, et on peut modéliser un lancer en disant que le résultat sera 1 avec probabilité $1/6$, 2 avec la même probabilité et ainsi de suite.

Un exemple moins classique : la description statistique d’un très grand tableau.

Imaginons qu’on interroge tou-te-s les français-es sur leur taille (arrondie au centimètre), et qu’on range ensuite les informations dans un grand tableau, avec trois colonnes (nom, prénom, taille), et 63 millions de lignes. Il est bien sûr impossible pour un humain de lire un tel tableau (seul un ordinateur pourrait le faire). Par contre, si on ne s’intéresse pas aux noms des français-es, on peut plutôt compter le nombre de français-se qui mesurent 150 cm, et obtenir la fréquence de la taille 150 cm en France.

En continuant ainsi pour toutes les valeurs de taille, on obtient ainsi un tableau de deux colonnes (taille, fréquence), et une centaine de lignes (pour des tailles comprises entre

90 cm et 230 cm ?), donc beaucoup plus lisibles.

C'est une description statistique. Le second tableau contient la **loi de probabilités** de la variable taille en France. On parle de **variable aléatoire**. La encore l'information est incomplète : on a volontairement négligé une partie des données, et on obtient aussi une autre sorte de hasard.

En effet, on ne peut plus choisir un-e français-e et lui associer sa taille. Mais on peut dire que si on avait choisi un-e français-e au hasard dans le tableau, on aurait obtenu la taille 150 cm avec telle probabilité, la taille 151 cm... et ainsi de suite.

Le cadre théorique : La théorie des probabilités. Celle-ci a pour objectif de formaliser cela. On l'utilise toujours dans les cas comme ci-dessus ou l'on ne dispose de suffisamment d'informations pour pouvoir déterminer de manière unique le résultat d'une expérience (lancer de dé, réponse à une question d'un sondage,...). Les résultats possibles s'appellent des **issues**, et l'ensemble de tous ceux-ci forme **l'espace des issues**.

A chacune de ces issues est associé une probabilité (un nombre compris entre 0 et 1 dans le cas où le nombre d'issues possibles est fini), qui donne la chance d'obtenir cette issue.

Plus précisément, cette probabilité est la fréquence de l'issue qu'on obtiendra si on répète infiniment de fois l'expérience en question (en faisant attention à ce qu'il n'y a pas de lien entre les résultats des diverses répétitions de l'expérience, on parle d'**indépendance**). C'est ici qu'apparaît le caractère théorique : il est bien sûr impossible de répéter une expérience un nombre infini de fois dans la pratique.

On appelle aussi **événement** un ensemble d'issues (exemples : obtenir un nombre pair en lançant un dé, avoir une taille supérieur à 170cm). La probabilité qu'on associe à un événement est la somme des probabilités de chaque issue de cet ensemble.

Intérêts de l'étude des probabilités. Cela permet d'évaluer les risques ou de mettre sur pieds des stratégies pour faire face aux aléas. La théorie des probabilités ne va pas permettre de prédire quelle issue va se réaliser, mais quelle chance a chaque issue de se réaliser.

1.2 Espace des issues.

Avant de calculer les probabilités d'évènements, il faut définir l'espace des issues de façon commode et complète. Cet espace comprendra toutes les issues possibles de l'expérience (avec résultat aléatoire) que l'on considère, même éventuellement celles qui ne nous intéressent pas, a priori. Dans chaque situation, l'espace des issues sera noté Ω (grand omega), alors que les issues seront notées ω (petit omega).

Exemple 1.1 On considère un dé à 6 faces, numérotées de 1 à 6. On suppose que le dé est équilibré, ce qui veut dire que les 6 faces ont la même chance de sortir. L'ensemble Ω des issues possibles d'un lancer est $\Omega = \{1, 2, 3, 4, 5, 6\}$, et on associe à chaque issue la probabilité $1/6$. On pense donc ici à une probabilité comme à une fonction P sur Ω , telle que

$$P(1) + \dots + P(6) = 1. \tag{1.1}$$

Quelle est la probabilité qu'un nombre pair sorte ? Il suffit de décomposer cet évènement A suivant les différentes issues qu'il contient : $A = \{2, 4, 6\}$, et

$$P(A) = P(2) + P(4) + P(6) = 1/2. \quad (1.2)$$

Exemple 1.2 On considère le même dé équilibré, sauf que sur la sixième face, le nombre 6 a été remplacé par 5. Il y a donc deux faces où 5 est inscrit. L'ensemble des issues est ici $\tilde{\Omega} = \{1, 2, 3, 4, 5\}$. Comme le dé est équilibré, les faces 1, 2, 3 et 4 ont chacune la probabilité $1/6$ de sortir, alors que $P(5) = 2/6$.

Exemple 1.3 Une personne I jette deux fois de suite le même dé équilibré. Elle note le résultat des deux lancers, et s'intéresse seulement à la somme (noté S) des deux dés.

L'espace des issues est

$$\Omega_I = \{(1, 1), (1, 2), (2, 1), \dots, (6, 6)\} = \{(i, j) : i, j \in \{1, \dots, 6\}\},$$

et $|\Omega| = 36$. Cet espace est d'autant plus commode que chaque couple a la même chance de sortir :

$$\forall i, j \in \{1, \dots, 6\}, \quad P(i, j) = \frac{1}{|\Omega|} = 1/36.$$

La somme des deux dés peut valoir n'importe quel nombre entier entre 2 et 12. Pour $i = 2, 3, \dots, 12$, la probabilité de l'évènement $S = i$ est ici moins immédiate. Par exemple si la somme des deux dés est 4, cela peut correspondre aux issues $\{(1, 3), (2, 2), (3, 1)\}$. Ainsi,

$$P(S = 4) = P(1, 3) + P(2, 2) + P(3, 1) = \frac{3}{36}.$$

Il faut faire la même chose pour trouver la probabilité associée à chaque évènement $S = i$.

Remarque importante : Si on s'intéresse qu'à la somme des deux dés S , on peut bien aussi oublier les deux lancer considérer que le résultat de l'expérience est uniquement

$$\Omega' = \{2, 3, 4, \dots, 11, 12\}.$$

Par exemple, si une personne lance les dés, calcule la somme, et donne le résultat à une seconde personne (qui n'a pas vu les lancer), Ω' est vraiment l'espace des issues pour cette seconde personne. Il est important de noter que le choix de l'espace des issues n'est pas unique ! Il y a une certaine liberté pour le choix de celui-ci, on peut décider d'oublier une partie de l'expérience (ou au contraire d'en rajouter). Ce qui doit vous guider est la facilité d'associer une probabilité à une issue.

Exemple 1.4 Pour illustrer notre propos, considérons le cas où une personne ne jette plus deux fois le même dé, mais jette en même temps deux dés identiques. L'information à laquelle elle a accès, n'est plus le couple (premier jet, deuxième jet), mais la paire {valeurs des deux dés}. Son espace des issues est donc

$$\tilde{\Omega} = \{\{i, j\}; i, j \in \{1, \dots, 6\}\}.$$

Noter que $\{1, 2\} = \{2, 1\}$, alors que $(1, 2) \neq (2, 1)$. La probabilité associée à chaque issue

$$\begin{cases} \tilde{P}(\{i, j\}) = P((i, j)) + P((j, i)) = \frac{2}{36} \text{ si } i \neq j; \\ \tilde{P}(\{i, i\}) = P((i, i)) = \frac{1}{36}. \end{cases}$$

Dans le calcul de ces probabilités, elle a en fait raisonné sur l'espace Ω , en considérant que l'issue $\{1, 2\}$ pouvait se réaliser de deux façons : le “premier” dé donne 1 et le “second” 2; et inversement. Tout en ne sachant pas dire quel est le “premier” dé, et quel est le “second”. Elle aurait donc tout aussi bien pu choisir comme espaces des issues l'espace Ω .

Exemple 1.5 On jette une pièce jusqu'à ce que Face sorte. Une façon de décrire l'expérience est de donner le nombre de lancers effectués. Ainsi, si 3 lancers ont été effectués, cela signifie que l'on a observé la suite PPF. Ainsi, l'espace Ω est ici l'ensemble \mathbb{N}^* des entiers naturels strictement positifs. De plus pour tout $k \in \mathbb{N}^*$,

$$P(k) = \underbrace{\frac{1}{2} \cdots \frac{1}{2}}_{k-1 \text{ fois P}} \underbrace{\frac{1}{2}}_{1 \text{ fois F}} = \left(\frac{1}{2}\right)^k.$$

On vérifie bien que $\sum_{k=1}^{\infty} p(k) = \sum_{k=1}^{\infty} \left(\frac{1}{2}\right)^k = \frac{1}{2} \sum_{k=0}^{\infty} \left(\frac{1}{2}\right)^k = \frac{1}{2} \frac{1}{1-\frac{1}{2}} = 1$.

1.3 Probabilités.

On considère dans un premier temps (et en particulier dans tout ce chapitre) des expériences avec un **nombre fini ou dénombrable d'issues**. On note Ω cet ensemble d'issues et $\mathcal{P}(\Omega)$ l'ensemble des évènements, c'est-à-dire l'ensemble des parties (ou sous-ensembles) de Ω . Si $\Omega = \{1, 2, 3\}$, alors

$$\mathcal{P}(\Omega) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 3\}, \{1, 2, 3\}\},$$

où \emptyset est l'ensemble vide.

Si A et B sont deux sous-ensembles de Ω , on note

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

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

$$A \setminus B = \{\omega : \omega \in A \text{ et } \omega \notin B\} \quad \text{lorsque } B \subset A,$$

$$A^c = \Omega \setminus A.$$

Noter que $(A^c)^c = A$ pour tout A . On interprète $A \cap B$ comme “ A et B se réalisent”; $A \cup B$ comme “ A ou B se réalisent”. On appelle A^c le complémentaire de A . A^c s'interprète comme “l'évènement A ne se réalise pas”.

Deux évènements A et B sont **disjoints** s'ils n'ont aucune issue en commun, c'est-à-dire que $A \cap B = \emptyset$. Par exemple, A et A^c sont disjoints, ainsi que \emptyset et A .

Exercice : Montrer que pour tous évènements A et B ,

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

Définition 1.6 Une **probabilité** P sur Ω est une fonction de $\mathcal{P}(\Omega)$ dans $[0, 1]$ telle que

- $P(\Omega) = 1$;
- Si $(A_i)_{i \in I}$ est une famille dénombrable d'évènements disjoints, (i.e. $A_i \cap A_j = \emptyset$ si $i \neq j$), alors

$$P(\cup_{i \in I} A_i) = \sum_{i \in I} P(A_i). \tag{1.3}$$

Le triplet $(\Omega, \mathcal{P}(\Omega), P)$ est appelé un **espace de probabilité**.

Pour souligner que l'ensemble Ω est au plus dénombrable, nous dirons aussi parfois que P est une **probabilité discrète**, et de même que $(\Omega, \mathcal{P}(\Omega), P)$ un **espace de probabilité discret**. Nous verrons dans la suite du cours une définition plus générale de ces notions.

De cette définition, découlent les propriétés suivantes :

Proposition 1.7 .

1. $P(\emptyset) = 0$.
2. Pour tout évènement A , $P(A) + P(A^c) = 1$.
3. Si A et B sont deux évènements tels que $A \subset B$, alors

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

4. Pour deux évènements quelconques A et B

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

5. Si $(A_n)_{n \geq 1}$ est une suite croissante d'évènements (i.e $A_n \subset A_{n+1}$ pour tout n), alors

$$P(\cup_{n \geq 1} A_n) = \lim_{n \rightarrow \infty} P(A_n).$$

6. Si $(B_n)_{n \geq 1}$ est une suite décroissante d'évènements (i.e $B_{n+1} \subset B_n$ pour tout n), alors

$$P(\cap_{n \geq 1} B_n) = \lim_{n \rightarrow \infty} P(B_n).$$

Preuve.

Preuve de 2. A et A^c sont deux évènements disjoints tels que $A \cup A^c = \Omega$. On a donc

$$1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c).$$

Preuve de 1. Il suffit de prendre $A = \emptyset$ dans 2.

Preuve de 3. Il suffit de remarquer que A et $B \setminus A$ sont deux évènements disjoints dont l'union est A .

Preuve de 4. Il suffit là encore de remarquer que A et $B \setminus (A \cap B)$ sont deux ensembles disjoints dont l'union est $A \cup B$, puis d'appliquer le résultat du point 3.

Preuve de 5. On pose $C_1 = A_1$ et pour $n \geq 2$, $C_n = A_n \setminus A_{n-1}$. Les évènements C_n sont deux à deux disjoints. De plus, on montre immédiatement par récurrence que pour tout $n \geq 1$, $\cup_{k=1}^n C_k = \cup_{k=1}^n A_k = A_n$, puis que $\cup_{n \geq 1} C_n = \cup_{n \geq 1} A_n$. Par conséquent,

$$\begin{aligned} P(\cup_{n \geq 1} A_n) &= P(\cup_{n \geq 1} C_n) = \sum_{n \geq 1} P(C_n) = \lim_{n \rightarrow \infty} \sum_{k=1}^n P(C_k) = \lim_{n \rightarrow \infty} P(\cup_{k=1}^n C_k) \\ &= \lim_{n \rightarrow \infty} P(A_n). \end{aligned}$$

Preuve de 6. On pose $A_n = B_n^c$. La suite $(A_n)_{n \geq 1}$ est alors une suite croissante et on peut appliquer le résultat du point 5 :

$$P(\cup_{n \geq 1} A_n) = \lim_{n \geq 1} P(A_n).$$

On utilise ensuite que pour tout ensemble A , $P(A^c) = 1 - P(A)$, et le point 6 découle alors de la formule $(\cup_{n \geq 1} A_n)^c = \cap_{n \geq 1} B_n$. ■

Remarque Ω étant un ensemble fini ou dénombrable, se donner une probabilité P sur l'ensemble des parties de Ω n'est rien d'autre que se donner la fonction sur $\Omega : \omega \in \Omega \mapsto p(\omega) = P(\{\omega\})$. En effet, pour tout événement A , on peut écrire A comme la réunion disjointe et au plus dénombrable des singletons des éléments qui le composent : $A = \cup_{\omega \in A} \{\omega\}$. Par la propriété (1.3), $P(A) = \sum_{\omega \in A} p(\omega)$.

Cette fonction vérifie les propriétés suivantes :

- $p : \Omega \mapsto [0, 1]$;
- $\sum_{\omega \in \Omega} p(\omega) = 1$.

Une telle fonction est une **distribution de probabilité**.

1.4 Variables aléatoires

1.4.1 Définition

Définition 1.8 Une fonction $X : \Omega \rightarrow \mathbb{R}$, est appelée **variable aléatoire** (*v.a.*).

De même que pour la notion de probabilité, pour souligner que Ω est au plus dénombrable, on parle aussi parfois de **variable aléatoire discrète**. Une définition plus générale sera donnée dans la suite du cours.

Ainsi, une variable aléatoire (discrète) n'est rien d'autre qu'une fonction tout à fait classique. Pourquoi donc une définition de plus ? En analyse, il est fondamental de savoir quel est le domaine de définition d'une fonction comme $\sin(x)$, ou $\log(x)$. On doit savoir où se trouve x : sur $[0, \pi[$, ou sur $]0, \infty[$, etc... En probabilité, ce qui est important ce sont les **valeurs** de la fonction, et les probabilités que l'on associe à chacune de ces valeurs. En revanche la connaissance précise de Ω importe peu, et dépend de notre façon de décrire le jeu, ou l'expérience que l'on modélise.

Convention. Les variables aléatoires sont en général notées par des lettres majuscules X, Y, \dots et les valeurs qu'elles prennent lorsque l'issue ω se réalise par $X(\omega), Y(\omega), \dots$

Exemple 1.9 Dans l'exemple 1.3 de deux lancers d'un dé, la somme des faces est une variable aléatoire :

$$\begin{aligned} X : \Omega_I &\rightarrow \mathbb{N} \\ (i, j) &\mapsto i + j. \end{aligned}$$

1.4.2 Loi d'une variable aléatoire

Soit Ω un espace d'issues (fini ou dénombrable), P une probabilité sur Ω , et X une variable aléatoire définie sur Ω . Soit $X(\Omega)$ l'ensemble des valeurs distinctes prises par X .

Par convention, si $x \in X(\Omega)$, on note

$$\begin{aligned} \{X = x\} &= \{\omega : X(\omega) = x\} & \text{et} & & P(X = x) &= P(\{\omega : X(\omega) = x\}), \\ \{X \leq x\} &= \{\omega : X(\omega) \leq x\} & \text{et} & & P(X \leq x) &= P(\{\omega : X(\omega) \leq x\}), \\ \{X < x\} &= \{\omega : X(\omega) < x\} & \text{et} & & P(X < x) &= P(\{\omega : X(\omega) < x\}), \end{aligned}$$

etc... De même, si A est un sous-ensemble de $X(\Omega)$,

$$\{X \in A\} = \{\omega : X(\omega) \in A\} \quad \text{et} \quad P(X \in A) = P(\{\omega : X(\omega) \in A\}).$$

En général si A est un sous-ensemble quelconque de \mathbb{R} , on définit $\{X \in A\}$ comme l'ensemble $\{X \in A \cap X(\Omega)\}$.

Définition 1.10 La loi de X est la probabilité P_X sur $X(\Omega)$, donnée par :

$$\forall A \subset X(\Omega), P_X(A) = P(X \in A). \quad (1.4)$$

Remarque Vérifier que P_X définie par (1.4), définit bien une probabilité sur $X(\Omega)$.

Remarque On rencontrera souvent des énoncés du type : “Soit X une variable aléatoire de loi ...”. Il faut comprendre ici que ce qui nous intéresse est la loi de X (les valeurs prises par X et les probabilités associées), et pas du tout l'espace Ω sur lequel est définie X . Dans beaucoup de situations, l'espace Ω ne sera même pas explicité.

Remarque En pratique lorsque l'on demande de donner la loi d'une variable aléatoire discrète, on pourra se contenter de donner sa distribution de probabilité, c'est-à-dire l'ensemble des probabilités $P(X = x)$, pour $x \in X(\Omega)$. En effet ces nombres déterminent complètement la probabilité P_X . À titre d'exercice donner la loi de probabilité de la variable X dans l'exemple 1.9.

Exemple 1.11 (Loi uniforme) La loi uniforme sur un ensemble fini E , est la loi d'une variable aléatoire X à valeurs dans E , qui prend chacune des valeurs de E avec la même probabilité $1/\text{Card}(E)$. Par exemple si $E = \{1, \dots, n\}$, X suit une loi uniforme sur E , si $P_X(k) = 1/n$ pour tout $k \in \{1, \dots, n\}$. On note alors $X \sim \mathcal{U}(\{1, \dots, n\})$, ou plus généralement $X \sim \mathcal{U}(E)$.

Exemple 1.12 (Loi de Bernoulli) La loi de Bernoulli de paramètre $p \in [0, 1]$ est la loi d'une variable aléatoire à valeurs dans $\{0, 1\}$, qui prend la valeur 1 avec probabilité p (et donc la valeur 0 avec probabilité $1 - p$). Si X suit cette loi, on note $X \sim \mathcal{B}(p)$.

Exemple 1.13 (Loi binomiale) La loi binomiale de paramètres $n \in \mathbb{N}$ et $p \in [0, 1]$ est la loi d'une variable aléatoire à valeurs dans $\{0, 1, \dots, n\}$, qui prend la valeur k avec probabilité $C_n^k p^k (1 - p)^{n-k}$ (où l'on rappelle que $C_n^k = \frac{n!}{k!(n-k)!}$). C'est la loi du nombre de gains lorsque l'on joue n fois de suite à un jeu où la probabilité de gagner est p . Si X suit cette loi, on note $X \sim \mathcal{B}(n; p)$.

Exemple 1.14 (Loi géométrique) La loi géométrique de paramètre $p \in]0, 1]$, est la loi d'une variable aléatoire à valeurs dans $\mathbb{N}^* = \{1, 2, \dots\}$, qui prend la valeur k avec probabilité $p(1 - p)^{k-1}$. C'est la loi du nombre de tentatives que l'on doit faire avant d'obtenir le premier gain, lorsque l'on joue successivement à un jeu où la probabilité de gagner est p . Si X suit cette loi, on note $X \sim \mathcal{G}(p)$.

Exemple 1.15 (Loi de Poisson) La loi de Poisson de paramètre $\lambda > 0$ est la loi d'une variable aléatoire à valeurs dans \mathbb{N} , qui prend la valeur k avec probabilité $e^{-\lambda} \lambda^k / k!$. Si X suit cette loi, on note $X \sim \mathcal{P}(\lambda)$. Cette loi est aussi appelée la loi des événements rares, car elle attribue une forte probabilité aux petites valeurs de k et une faible probabilité aux grandes valeurs de k (le terme $\lambda^k / k!$ convergeant très vite vers 0 lorsque k grandit). C'est par exemple le type de loi qui peut être utilisé pour modéliser le nombre d'enfants d'un couple choisi au hasard dans une population donnée, ou le nombre d'accidents mortels sur la route un jour donné.

1.4.3 Fonction de répartition

Définition 1.16 La fonction de répartition d'une variable aléatoire X est la fonction F_X définie par :

$$F_X : \mathbb{R} \rightarrow [0, 1] \\ x \mapsto P(X \leq x)$$

Proposition 1.17 Soit F_X la fonction de répartition d'une variable aléatoire X .

1. Si $x \leq y$, $F_X(x) \leq F_X(y)$. F_X est une fonction croissante.
2. F_X est continue à droite en tout point.
3. $\lim_{x \rightarrow -\infty} F_X(x) = 0$.
4. $\lim_{x \rightarrow +\infty} F_X(x) = 1$.
5. $\forall x \in \mathbb{R}$, $P(X = x) = F_X(x) - \lim_{y \rightarrow x, y < x} F_X(y)$.

Preuve.

Preuve de 1. Si $x \leq y$, $\{X \leq x\} \subset \{X \leq y\}$. Par conséquent $F_X(x) = P[X \leq x] \leq P[X \leq y] = F_X(y)$.

Preuve de 2. Pour tout $x \in \mathbb{R}$, $\{X \leq x\} = \bigcap_{n \in \mathbb{N}} \{X \leq x + \frac{1}{n}\}$. De plus cette intersection est une intersection décroissante d'évènements. En utilisant la propriété 6 de la proposition 1.7, on a donc

$$F_X(x) = P(X \leq x) = \lim_{n \rightarrow \infty} P(X \leq x + \frac{1}{n}) = \lim_{n \rightarrow \infty} F_X(x + \frac{1}{n}),$$

ce qui prouve la continuité à droite de F_X (en utilisant que F_X est croissante).

De même propriété 5 de la proposition 1.7 montre que la limite $\lim_{y \rightarrow x, y < x} F_X(y)$ existe et vaut $P(X < x)$.

Preuve de 3. Soit $(y_n)_{n \in \mathbb{N}}$ une suite décroissant vers $-\infty$. On a $\emptyset = \bigcap_{n \in \mathbb{N}} \{X \leq y_n\}$, et cette intersection est une intersection décroissante d'évènements. On a donc

$$0 = P(\emptyset) = \lim_{n \rightarrow \infty} P(X \leq y_n) = \lim_{x \rightarrow -\infty} F_X(x).$$

Preuve de 4. Soit $(y_n)_{n \in \mathbb{N}}$ une suite croissant vers $+\infty$. On a $\Omega = \bigcup_{n \in \mathbb{N}} \{X \leq y_n\}$, et cette union est une union croissante d'évènements. On a donc

$$1 = P(\Omega) = \lim_{n \rightarrow \infty} P(X \leq y_n) = \lim_{x \rightarrow +\infty} F_X(x).$$

Preuve de 5. Soit $x \in \mathbb{R}$ fixé. On a $\{X = x\} = \{X \leq x\} \setminus \{X < x\}$. On a donc

$$P(X = x) = P(X \leq x) - P(X < x),$$

et on conclut avec la propriété 1. ■

Proposition 1.18 La fonction de répartition d'une variable aléatoire détermine sa loi.

Preuve. Il s'agit de voir qu'à partir de la donnée de la fonction F_X , on est capable de retrouver l'ensemble des valeurs prises par X , et leurs probabilités associées. Or d'après la proposition précédente, les valeurs prises par X sont les points où F_X est discontinue, et les probabilités associées sont les amplitudes des sauts en ces points de discontinuité. ■

Remarque En fait dans le cas où l'on peut ordonner l'ensemble des valeurs prises par X en une suite croissante $(\dots, x_{-1}, x_0, x_1, \dots)$, les résultats précédents montrent que la fonction F_X est constante sur chaque intervalle $[x_n, x_{n+1}[$.

1.4.4 Fonction indicatrice.

La **fonction indicatrice** $\mathbb{1}_A$ d'un évènement $A \subset \Omega$, est la variable aléatoire

$$\begin{aligned} \mathbb{1}_A : \Omega &\rightarrow \{0, 1\} \\ \omega &\mapsto \begin{cases} 1 & \text{si } \omega \in A \\ 0 & \text{si } \omega \notin A \end{cases} \end{aligned}$$

La loi de cette variable est donc la **loi de Bernoulli** de paramètre $p = P(A)$. Nous allons voir que toute variable aléatoire (discète) peut s'écrire comme une combinaison linéaire de fonctions indicatrices, d'où leur importance.

On rappelle tout d'abord qu'une **partition** de Ω est une suite de sous-ensembles $(A_i)_{i \in I}$ disjoints deux à deux, de réunion Ω . En d'autres termes,

$$\text{Pour tout } i \text{ différent de } j, A_i \cap A_j = \emptyset \quad \text{et} \quad \cup_{i \in I} A_i = \Omega.$$

La proposition suivante est élémentaire.

- Proposition 1.19**
1. $\mathbb{1}_{A \cap B} = \mathbb{1}_A \mathbb{1}_B$.
 2. Si $A \cap B = \emptyset$, alors $\mathbb{1}_{A \cup B} = \mathbb{1}_A + \mathbb{1}_B$.
 3. $\mathbb{1}_{A^c} = 1 - \mathbb{1}_A$.
 4. Si $(A_i)_{i \in I}$ est une partition de Ω , alors $1 = \sum_{i \in I} \mathbb{1}_{A_i}$.

Preuve. Nous ne démontrons que le point 4. Soit ω un élément fixé de Ω . $(A_i)_{i \in I}$ étant une partition, ω appartient à un et un seul des ensembles A_i . Autrement dit, il existe un unique i tel que $\mathbb{1}_{A_i}(\omega) = 1$. Ainsi, dans la somme $\sum_{i \in I} \mathbb{1}_{A_i}(\omega) = 1$, un seul des termes est non nul, et vaut 1. ■

Définition 1.20 Soit $(A_i)_{i \in I}$ une partition de Ω . On dit que X se **décompose sur la partition** $(A_i)_{i \in I}$ si X peut s'écrire sous la forme

$$X = \sum_{i \in I} x_i \mathbb{1}_{A_i}; \tag{1.5}$$

ce qui signifie que X est constante sur les évènements A_i et prend la valeur x_i sur A_i .

Exemple 1.21 Dans un alphabet de 26 lettres, on forme un mot de 15 lettres en choisissant chaque lettre indépendamment des autres et avec la même probabilité. Pour chaque entier $i \leq 26$, on note B_i , l'évènement selon lequel la lettre i n'apparaît pas dans le mot, et $A_i = B_i^c$, l'évènement "la lettre i apparaît au moins une fois". Alors la variables aléatoire X donnant le nombre de lettres distinctes qui apparaissent dans un mot ω se décompose comme

$$X = \sum_{i=1}^{26} \mathbb{1}_{A_i}(\omega).$$

A toute variable X , on peut associer une partition de Ω de la façon suivante : soit $\{x_i\}_{i \in I}$, l'ensemble des valeurs distinctes prises par X . Posons

$$A_i = \{\omega : X(\omega) = x_i\}.$$

Les A_i forment une partition de Ω . En effet,

$$A_i \cap A_j = \{\omega : X(\omega) = x_i \text{ et } X(\omega) = x_j\} = \emptyset, \text{ si } i \neq j,$$

$$\text{et } \cup_i A_i = \{\omega : \exists i \text{ tel que } X(\omega) = x_i\} = \Omega.$$

De plus, par définition des A_i , on a $X = \sum_{i \in I} x_i \mathbb{1}_{A_i}$.

En revanche, l'écriture (1.5) n'est pas unique. Par exemple, si A, B, C forment une partition de Ω , alors

$$X = 3 \mathbb{1}_{A \cup B} + \mathbb{1}_C = 3 \mathbb{1}_A + 3 \mathbb{1}_B + \mathbb{1}_C.$$

Enfin, on peut noter que si X se décompose sur une partition $(A_i)_{i \in I}$ ($X = \sum_i x_i \mathbb{1}_{A_i}$), toute fonction de X se décompose sur la même partition. Si $f : \mathbb{R} \rightarrow \mathbb{R}$, alors

$$f(X) = \sum_i f(x_i) \mathbb{1}_{A_i}. \tag{1.6}$$

Notons qu'il se peut que les valeurs $\{f(x_i)\}_i$ ne soient pas toutes distinctes.

Chapitre 2

Espérance, variance, écart type.

2.1 Espérance.

2.1.1 Définitions.

Soit Ω un espace d'issues fini ou dénombrable, P une probabilité sur Ω , et $A \subset \Omega$. Imaginons que chaque fois qu'une issue ω de A se réalise, on gagne 1 euro, et qu'on ne gagne rien sinon. Par exemple, on gagne un euro si lorsqu'on jette deux dés, la somme est paire. Notre gain est donc $\mathbb{1}_A(\omega)$. Combien gagne-t-on en moyenne ? Intuitivement, on va gagner en moyenne $P(A)$. Plus généralement, supposons que $\{A_i, i \in I\}$ est une partition de Ω , et que l'on gagne x_i euros si $\omega \in A_i$. Par exemple, on gagne $x_1 = 1$ euro, si la somme des dés est paire ; on perd 2 euros (i.e. $x_2 = -2$) si elle est impaire. Le gain est ici

$$X = x_1 \mathbb{1}_{A_1} + x_2 \mathbb{1}_{A_2} + \cdots = \sum_{i \in I} x_i \mathbb{1}_{A_i}.$$

Combien gagne-t-on en moyenne ? Intuitivement, on gagne x_i euros avec probabilité $P(A_i)$ pour chaque i de 1 à n . Ainsi, le gain moyen noté $E[X]$, est

$$\begin{aligned} E[X] &= x_1 P(A_1) + x_2 P(A_2) + \cdots = \sum_{i \in I} x_i P(A_i). \\ &= \sum_{i \in I} x_i \left(\sum_{\omega \in A_i} P(\omega) \right) \\ &= \sum_{i \in I} \sum_{\omega \in A_i} X(\omega) P(\omega), \text{ puisque } X(\omega) = x_i \text{ si } \omega \in A_i, \\ &= \sum_{\omega \in \Omega} X(\omega) P(\omega) \text{ car les } A_i \text{ forment une partition.} \end{aligned} \tag{2.1}$$

Cela nous amène à la définition suivante :

Définition 2.1 Soit Ω un espace fini ou dénombrable, P une probabilité sur Ω , et X une variable aléatoire sur Ω . On appelle **espérance** de X (ou **moyenne** de X) la quantité

$$E[X] = \sum_{\omega \in \Omega} X(\omega) P(\omega). \tag{2.2}$$

Remarque L'espérance de X n'est bien définie que si la série de terme général $X(\omega)P(\omega)$ est absolument convergente, soit si $\sum_{\omega \in \Omega} |X(\omega)P(\omega)|$ est fini. Dans toute la suite, lorsqu'on parlera de l'espérance d'une variable, on sous-entendra toujours que cette condition est vérifiée.

Remarque $E[X]$ est un **barycentre des valeurs possibles pour X** . En ce sens, $E[X]$ est un paramètre de **position**.

En remontant la suite d'égalités de (2.1), on démontre à partir de la définition, la proposition

Proposition 2.2 Si $(A_i)_{i \in I}$ est une partition de Ω , et si X se décompose sur cette partition, i.e. si X peut s'écrire $X = \sum_{i \in I} x_i \mathbb{1}_{A_i}$, alors

$$E[X] = \sum_{i \in I} x_i P(A_i). \quad (2.3)$$

Remarque On a ainsi $E[\mathbb{1}_A] = P(A)$.

Remarque Il aurait été plus naturel de prendre l'identité (2.3) comme définition de l'espérance. Toutefois, cette identité dépend a priori du choix d'une partition sur laquelle on peut décomposer X . Un tel choix n'est évidemment pas unique. L'écriture (2.2) montre que $E[X]$ ne dépend pas du choix de la décomposition de X . C'est le principal intérêt de cette formule, qui sert peu souvent dans les calculs où on lui préférera (2.3).

Remarque Si $(A_i)_{i \in I}$ est une partition de Ω , et si X se décompose sur cette partition ($X = \sum_{i \in I} x_i \mathbb{1}_{A_i}$), il en est de même pour toute fonction $f(X)$ de X ,

$$f(X) = \sum_{i \in I} f(x_i) \mathbb{1}_{A_i}.$$

Ainsi, $E[f(X)] = \sum_{i \in I} f(x_i) P(A_i)$.

Remarque Pour calculer $E[X]$, il suffit de connaître la loi de X , i.e. les valeurs distinctes $(x_i)_{i \in I}$ prises par X et les probabilités $P(X = x_i)$. En effet, si on note $A_i = \{X = x_i\}$, on a $X = \sum_{i \in I} x_i \mathbb{1}_{A_i}$, et

$$E[X] = \sum_{i \in I} x_i P(X = x_i).$$

De même, pour toute fonction f , $f(X) = \sum_{i \in I} f(x_i) \mathbb{1}_{A_i}$, et

$$E[f(X)] = \sum_{i \in I} f(x_i) P(X = x_i).$$

Exemple 2.3 On jette un dé équilibré. Si un nombre pair sort, on gagne 2 euros. Si un nombre impair sort, on perd 3 euros. Combien gagne-t-on en moyenne ? On modélise notre jeu par

$$\Omega = \{1, 2, 3, 4, 5, 6\} \quad \text{et} \quad \forall i \in \{1, \dots, 6\}, p(i) = 1/6.$$

Le gain X est la variable définie par

$$\begin{cases} X(\omega) = -3 & \text{si } \omega \in A = \{1, 3, 5\}. \\ X(\omega) = 2 & \text{si } \omega \in B = \{2, 4, 6\}. \end{cases}$$

Ainsi $X = -3 \mathbb{I}_A + 2 \mathbb{I}_B$, et $E[X] = -3P(A) + 2P(B) = -3\frac{1}{2} + 2\frac{1}{2} = -\frac{1}{2}$.

2.1.2 Propriétés de l'espérance.

Proposition 2.4 (Linéarité de l'espérance).

Soient X et Y deux variables aléatoires et $\lambda \in \mathbb{R}$. Alors,

1. $E[\lambda X] = \lambda E[X]$.
2. $E[X + Y] = E[X] + E[Y]$.

Preuve.

1) Par définition,

$$E[\lambda X] = \sum_{\omega \in \Omega} \lambda X(\omega) P(\omega) = \lambda \sum_{\omega \in \Omega} X(\omega) P(\omega) = \lambda E[X].$$

2)

$$\begin{aligned} E[X + Y] &= \sum_{\omega \in \Omega} (X(\omega) + Y(\omega)) P(\omega) = \sum_{\omega \in \Omega} X(\omega) P(\omega) + \sum_{\omega \in \Omega} Y(\omega) P(\omega) \\ &= E[X] + E[Y]. \end{aligned}$$

Nous énonçons sans démonstration certaines propriétés simples de l'espérance. Il est important de les vérifier seul! ■

Proposition 2.5 *Soient X, Y deux variables aléatoires sur (Ω, P) , et $\lambda \in \mathbb{R}$.*

1. $E[\lambda] = \lambda$.
2. $X \geq 0 \implies E[X] \geq 0$.
3. $X \geq Y \implies E[X] \geq E[Y]$.
4. $|E[X]| \leq E[|X|]$.

Remarque Une variable aléatoire dont l'espérance est nulle est dite **centrée**.

2.2 Variance, covariance, et écart type.

Définition 2.6 *La variance d'une variable aléatoire X est le nombre positif donné par*

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

Son écart type est le nombre

$$\sigma(X) = \sqrt{\text{Var}(X)}.$$

Remarque $\text{Var}(X) = \sum_{\omega \in \Omega} (X(\omega) - E[X])^2 P(\omega)$ est le barycentre des carrés des distances des valeurs prises par X à la valeur moyenne de X . C'est donc un paramètre de **dispersion**.

Remarque La variance d'une variable aléatoire X est bien définie lorsque la série de terme général $X(\omega)^2 P(\omega)$ est finie (noter en particulier que si cette condition est vérifiée, alors en utilisant que $\sum_{\omega \in \Omega} P(\omega) = 1$ et l'inégalité de Cauchy-Schwarz¹, on en déduit que la série de terme général $X(\omega)P(\omega)$ est absolument convergente, et donc que l'espérance de X est également bien définie). Cette hypothèse sera toujours faite implicitement lorsque l'on parlera de la variance d'une variable aléatoire.

Proposition 2.7 (Propriétés de la variance)

Soit X une variable aléatoire et $\lambda \in \mathbb{R}$.

1. $\text{Var}(\lambda X) = \lambda^2 \text{Var}(X)$.
2. $\text{Var}(X + \lambda) = \text{Var}(X)$.
3. $\text{Var}(\lambda) = 0$.
4. $\text{Var}(X) = 0$ ssi $\forall \omega \in \Omega$ tel que $P(\omega) > 0$, $X(\omega) = E[X]$.
5. $\text{Var}(X) = E[X^2] - (E[X])^2$.

Remarque La propriété 4. signifie que $\text{var}(X) = 0$ ssi X est constante sur les issues possibles.

Preuve.

$$\begin{aligned} 1. \text{Var}(\lambda X) &= E \left[(\lambda X - E[\lambda X])^2 \right] \\ &= E \left[\lambda^2 (X - E[X])^2 \right] \text{ par linéarité de l'espérance,} \\ &= \lambda^2 \text{Var}(X) \text{ par linéarité de l'espérance.} \end{aligned}$$

$$2. \text{Var}(X + \lambda) = E \left[(X + \lambda - E[X + \lambda])^2 \right] = E \left[(X + \lambda - E[X] - \lambda)^2 \right] = \text{Var}(X).$$

$$3. \text{Var}(\lambda) = E \left[(\lambda - E[\lambda])^2 \right] = E \left[(\lambda - \lambda)^2 \right] = E[0] = 0.$$

4. Comme $\text{Var}(X) = \sum_{\omega \in \Omega} (X(\omega) - E[X])^2 P(\omega)$ est une somme de termes positifs, pour que $\text{Var}(X) = 0$, il faut et il suffit que tous les termes de la somme soient nuls.

5. On développe le carré, $(X - E[X])^2 = X^2 - 2X.E[X] + (E[X])^2$. Par linéarité de l'espérance,

$$\text{Var}(X) = E[X^2] - 2E[X]E[X] + E[X]^2 = E[X^2] - E[X]^2.$$

■

Définition 2.8 La **covariance** de deux variables aléatoires X et Y est le nombre

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

1. cette inégalité nous dit que si $(u_n)_{n \geq 0}$ et $(v_n)_{n \geq 0}$ sont deux suites réelles positives quelconques, alors $(\sum_n u_n v_n)^2 \leq (\sum_n u_n^2)(\sum_n v_n^2)$. Ici on l'applique avec $u_n = |X(\omega)|\sqrt{P(\omega)}$ et $v_n = \sqrt{P(\omega)}$.

Remarque Ici encore la covariance de X et Y n'est bien définie que si la série de terme général $X(\omega)Y(\omega)P(\omega)$ est absolument convergente, hypothèse qui sera toujours faite implicitement. En fait on supposera même que les variances de X et Y sont bien définies, ce qui d'après l'inégalité de Cauchy-Schwarz est une hypothèse plus forte. On a alors

$$|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}.$$

Proposition 2.9 Soient deux variables aléatoires X et Y . On a

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

Preuve. Il suffit d'écrire la définition de la variance et de développer. Puis en utilisant la linéarité de l'espérance, on obtient :

$$\begin{aligned} \text{Var}(X + Y) &= E[(X + Y - E[X + Y])^2] \\ &= E[(X - E[X])^2 + (Y - E[Y])^2 + 2E[(X - E[X])(Y - E[Y])] \\ &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y). \end{aligned}$$

■

2.3 Espérance et variance de quelques lois classiques.

Proposition 2.10 Soit X une variable de loi de Bernoulli de paramètre p . Alors

$$E[X] = p \quad \text{et} \quad \text{Var}(X) = p(1 - p).$$

Preuve. Par définition, $X \in \{0, 1\}$ et $P(X = 1) = p = 1 - P(X = 0)$. Donc

$$E[X] = 1 P(X = 1) + 0 P(X = 0) = p.$$

De plus, $X^2 = 0 \mathbb{1}_{\{X=0\}} + 1 \mathbb{1}_{\{X=1\}}$. Par conséquent,

$$\begin{aligned} E[X^2] &= 0 P(X = 0) + 1 P(X = 1) = p, \\ \text{Var}(X) &= E[X^2] - (E[X])^2 = p - p^2 = p(1 - p). \end{aligned}$$

■

Proposition 2.11 Soit S_n une variable de loi binomiale de paramètre n et p .

$$E[S_n] = np, \quad \text{Var}(S_n) = np(1 - p).$$

Preuve. On utilise la décomposition $S_n = \sum_{i=0}^n i \mathbb{1}_{S_n=i}$ et les valeurs $P(S_n = i) = C_n^i p^i (1 - p)^{n-i}$. En utilisant ensuite l'identité $i C_n^i = n C_{n-1}^{i-1}$, valable pour $1 \leq i \leq n$, on en déduit

$$\begin{aligned} E[S_n] &= \sum_{i=0}^n i C_n^i p^i (1 - p)^{n-i} \\ &= np \sum_{i=1}^n C_{n-1}^{i-1} p^{i-1} (1 - p)^{n-i} \\ &= np \sum_{i=0}^{n-1} C_{n-1}^i p^i (1 - p)^{n-1-i} \\ &= np(p + 1 - p)^{n-1} \\ &= np. \end{aligned}$$

Puis en utilisant que $S_n^2 = \sum_{i=1}^n i^2 \mathbb{1}_{S_n=i}$, on trouve de même

$$\begin{aligned}
E[S_n^2] &= \sum_{i=0}^n i^2 C_n^i p^i (1-p)^{n-i} \\
&= np \sum_{i=1}^n i C_{n-1}^{i-1} p^{i-1} (1-p)^{n-i} \\
&= np \left(\sum_{i=1}^n (i-1) C_{n-1}^{i-1} p^{i-1} (1-p)^{n-i} + \sum_{i=1}^n C_{n-1}^{i-1} p^{i-1} (1-p)^{n-i} \right) \\
&= np((n-1)p + 1) \\
&= (np)^2 + np(1-p).
\end{aligned}$$

On trouve alors $\text{Var}(S_n) = E[S_n^2] - E[S_n]^2 = np(1-p)$. ■

Proposition 2.12 *Soit X une variable de loi géométrique de paramètre $p \in]0, 1]$.*

$$E[X] = \frac{1}{p}, \quad \text{Var}(X) = \frac{1-p}{p^2}.$$

Preuve.

$$\begin{aligned}
E[X] &= \sum_{k=1}^{\infty} k P(X=k) = p \sum_{k=1}^{\infty} k (1-p)^{k-1} = -p \frac{d}{dp} \left(\sum_{k=0}^{\infty} (1-p)^k \right) \\
&= -p \frac{d}{dp} \left(\frac{1}{p} \right) = \frac{p}{p^2} = \frac{1}{p}.
\end{aligned}$$

$$\begin{aligned}
E[X^2] &= \sum_{k=1}^{\infty} k^2 P(X=k) = p \sum_{k=1}^{\infty} k^2 (1-p)^{k-1} \\
&= p \sum_{k=1}^{\infty} k(k-1) (1-p)^{k-1} + p \sum_{k=1}^{\infty} k (1-p)^{k-1} \\
&= p(1-p) \sum_{k=2}^{\infty} k(k-1) (1-p)^{k-2} + E[X] \\
&= p(1-p) \frac{d^2}{dp^2} \left(\sum_{k=0}^{\infty} (1-p)^k \right) + E[X] = p(1-p) \frac{d^2}{dp^2} \left(\frac{1}{p} \right) + \frac{1}{p} \\
&= -p(1-p) \frac{d}{dp} \left(\frac{1}{p^2} \right) + \frac{1}{p} = \frac{2p(1-p)}{p^3} + \frac{1}{p} \\
&= \frac{2}{p^2} - \frac{1}{p}.
\end{aligned}$$

Ainsi, $\text{Var}(X) = E[X^2] - E[X]^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}$. ■

Proposition 2.13 *Soit X une variable de loi de Poisson de paramètre $\lambda > 0$. Alors*

$$E[X] = \lambda, \quad \text{Var}(X) = \lambda.$$

Preuve.

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} kP(X = k) = \sum_{k=0}^{\infty} ke^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} \\ &= e^{-\lambda} \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = e^{-\lambda} \lambda e^{\lambda} = \lambda. \end{aligned}$$

$$\begin{aligned} E[X^2] &= \sum_{k=0}^{\infty} k^2 P(X = k) = \sum_{k=0}^{\infty} k^2 e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=1}^{\infty} k \frac{\lambda^k}{(k-1)!} \\ &= e^{-\lambda} \lambda \sum_{k=1}^{\infty} (k-1) \frac{\lambda^{k-1}}{(k-1)!} + e^{-\lambda} \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = e^{-\lambda} \lambda \sum_{k=2}^{\infty} \frac{\lambda^{k-1}}{(k-2)!} + \lambda \\ &= e^{-\lambda} \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \lambda = \lambda^2 + \lambda. \end{aligned}$$

$$\text{Et } \text{Var}(X) = E[X^2] - E[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda. \quad \blacksquare$$

Chapitre 3

Indépendance, Probabilités Conditionnelles.

3.1 Indépendance.

La notion intuitive d'indépendance s'illustre avec les exemples suivants :

Exemple 3.1 On lance deux fois un dé, et l'on demande la probabilité que le deuxième jet donne 3, si le premier a donné 5? La réponse immédiate est qu'on a 1 chance sur 6 que le deuxième jet donne 3, quelle que soit en fait la valeur obtenue lors du premier jet. *Les deux lancers sont indépendants.* On peut noter que dans l'exemple 1.3 on a traduit cette indépendance en disant que chaque couple (i, j) de l'espace Ω_I avait une probabilité $\frac{1}{36} = \frac{1}{6} \times \frac{1}{6}$ de se réaliser; i.e. *la probabilité du couple (i, j) est le produit des probabilités de chaque jet.*

Exemple 3.2 Imaginons 2 urnes I et II :

- L'urne I contient 4 boules noires et 1 boule blanche.
- L'urne II contient 1 boule noire, et 99 boules blanches.

Le jeu consiste à choisir une urne "au hasard", et dans cette urne, une boule "au hasard". Par "au hasard", on entend que tous les choix possibles ont la même chance d'être réalisés. Est-ce que le numéro de l'urne choisie, et la couleur de la boule tirée sont indépendants? Evidemment que non, puisque l'urne que je choisis, conditionne les probabilités des couleurs de la boule tirée.

Nous allons tenter de rendre ces intuitions rigoureuses à l'aide d'une définition adaptée.

Soit Ω un ensemble d'issues (fini ou dénombrable), et P une probabilité sur Ω .

Définition 3.3 Deux évènements A et B sont **indépendants** si

$$P(A \cap B) = P(A)P(B). \quad (3.1)$$

Remarque Attention à ne pas confondre l'indépendance de deux évènements A et B avec la condition $A \cap B = \emptyset$. Cette dernière condition implique $P(A \cap B) = 0$ (et $P(A \cup B) = P(A) + P(B)$). En particulier à moins que $P(A)$ ou $P(B)$ ne soit nul, elle implique au contraire que A et B ne sont pas indépendants!

Définition 3.4 Deux variables aléatoires $X : \Omega \mapsto \{x_i, i \in I\}$ et $Y : \Omega \mapsto \{y_j, j \in J\}$ sont **indépendantes** si pour tous i et j ,

$$P(\{X = x_i\} \cap \{Y = y_j\}) = P(X = x_i)P(Y = y_j).$$

On généralise cette définition à un nombre fini ou dénombrable d'évènements ou de variables aléatoires :

Définition 3.5 Les évènements $A_i, i \in I$, sont **indépendants** si pour tout sous-ensemble fini d'indices $J \subset I$,

$$P\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} P(A_j).$$

Par exemple, pour montrer que trois évènements A, B et C sont indépendants, il faut vérifier que

$$P(A \cap B \cap C) = P(A)P(B)P(C), \quad P(A \cap B) = P(A)P(B),$$

$$P(A \cap C) = P(A)P(C), \quad \text{et} \quad P(C \cap B) = P(C)P(B).$$

Définition 3.6 Les variables aléatoires $X_i, i \in I$, sont **indépendantes**, si pour tout sous-ensemble fini d'indices $J \subset I$ et tous réels x_j appartenant aux valeurs possibles de $X_j, j \in J$,

$$P\left(\bigcap_{j \in J} \{X_j = x_j\}\right) = \prod_{j \in J} P(X_j = x_j).$$

Remarque On notera en particulier que l'indépendance d'une suite de variables aléatoires (ou évènements) est une propriété plus forte que l'indépendance deux à deux des mêmes variables aléatoires. Ceci se voit sur l'exemple suivant. Soit X_1, X_2 et X_3 trois v.a. indépendantes, telles que pour $i = 1, 2, 3$, on ait $P(X_i = 1) = P(X_i = -1) = 1/2$. Posons $Y_1 = X_1X_2, Y_2 = X_2X_3$ et $Y_3 = X_3X_1$. Les trois v.a. Y_1, Y_2, Y_3 ne sont pas indépendantes, puisque $Y_3 = Y_1Y_2$ donc connaissant Y_1 et Y_2 , on connaît Y_3 et l'indépendance globale des trois v.a. est mise en défaut (exercice). Par contre, ces trois v.a. sont deux à deux indépendantes (exercice. Notez qu'il suffit en fait de vérifier que Y_1 et Y_2 sont indépendantes, les autres vérifications étant identiques).

Proposition 3.7 Soient X_1, \dots, X_n, n variables aléatoires indépendantes et soit $f_1, \dots, f_n : \mathbb{R} \rightarrow \mathbb{R}$, des fonctions quelconques. Alors $f_1(X_1), \dots, f_n(X_n)$ sont indépendantes.

Preuve. Nous montrons le résultat lorsque $n = 2$. Le cas général s'en déduit immédiatement par récurrence. Soient donc X et Y deux variables aléatoires indépendantes, et f, g deux fonctions quelconques. Soient $\{x_i; i \in I\}$ les valeurs possibles pour X , et $\{y_j; j \in J\}$ les valeurs possibles pour Y . Les valeurs possibles de $f(X)$ sont $\{f(x_i); i \in I\}$. Parmi ces valeurs, certaines peuvent être identiques. Notons alors $\{f_k; k \in K\}$ les valeurs distinctes prises par $f(X)$. De même, notons $\{g_l; l \in L\}$ les valeurs distinctes prises par $g(Y)$. Pour k et l fixés, on a

$$\{f(X) = f_k\} = \bigcup_{i: f(x_i)=f_k} \{X = x_i\},$$

$$\{g(Y) = g_l\} = \bigcup_{j:g(y_j)=g_l} \{Y = y_j\}.$$

Si on pose $A_i = \{X = x_i\}$ et $B_j = \{Y = y_j\}$, on voit que les $\{A_i \cap B_j, i \in I, j \in J\}$ sont disjoints, et donc que

$$\begin{aligned} & P(\{f(X) = f_k\} \cap \{g(Y) = g_l\}) \\ &= P\left(\left(\bigcup_{i:f(x_i)=f_k} A_i\right) \cap \left(\bigcup_{j:g(y_j)=g_l} B_j\right)\right) \\ &= P\left(\bigcup_{i:f(x_i)=f_k} \bigcup_{j:g(y_j)=g_l} (A_i \cap B_j)\right) \\ &= \sum_{i:f(x_i)=f_k} \sum_{j:g(y_j)=g_l} P(A_i \cap B_j) \\ &= \sum_{i:f(x_i)=f_k} \sum_{j:g(y_j)=g_l} P(A_i)P(B_j) \text{ par indépendance de } X \text{ et } Y, \\ &= \left(\sum_{i:f(x_i)=f_k} P(A_i)\right) \left(\sum_{j:g(y_j)=g_l} P(B_j)\right) \\ &= P\left(\bigcup_{i:f(x_i)=f_k} A_i\right) P\left(\bigcup_{j:g(y_j)=g_l} B_j\right) \\ &= P(f(X) = f_k)P(g(Y) = g_l). \end{aligned}$$

■

Proposition 3.8 Soient deux variables aléatoires X et Y **indépendantes**. Alors

$$E[XY] = E[X]E[Y].$$

En particulier $\text{Cov}(X, Y) = 0$, et donc

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

Preuve. Soient $(x_i)_{i \in I}$ les valeurs distinctes prise par X , et $(y_j)_{j \in J}$ les valeurs distinctes prises par Y . Posons $A_i = \{X = x_i\}$ et $B_j = \{Y = y_j\}$ les partitions respectivement associées à X et Y . Les $A_i \cap B_j$ ($i \in I, j \in J$) forment une partition de Ω , et on peut écrire

$$X = \sum_{i \in I} \sum_{j \in J} x_i \mathbb{1}_{A_i \cap B_j}, \quad Y = \sum_{i \in I} \sum_{j \in J} y_j \mathbb{1}_{A_i \cap B_j}, \quad \text{et } XY = \sum_{i \in I} \sum_{j \in J} x_i y_j \mathbb{1}_{A_i \cap B_j}.$$

Donc,

$$E[XY] = \sum_{i,j} x_i y_j P(A_i \cap B_j).$$

X et Y étant indépendantes, $P(A_i \cap B_j) = P(A_i)P(B_j)$. Ainsi,

$$\begin{aligned} E[XY] &= \sum_{i,j} x_i y_j P(A_i) P(B_j) \\ &= \left(\sum_{i \in I} x_i P(A_i) \right) \left(\sum_{j \in J} y_j P(B_j) \right) \\ &= E[X] E[Y]. \end{aligned}$$

Le dernier point de la proposition découle alors de la proposition 2.9. ■

Remarque Attention, la réciproque du dernier point de la proposition précédente est fautive : on peut avoir $\text{Cov}(X, Y) = 0$, sans que X et Y soient indépendantes. Par exemple si X est une variable de Bernoulli de paramètre $1/2$ et ϵ une variable indépendante dont la loi est donnée par $P(\epsilon = 1) = P(\epsilon = -1) = 1/2$, alors en posant $Y = \epsilon X$, on voit que $\text{Cov}(X, Y) = 0$, mais X et Y ne sont pas indépendantes.

Remarque (Construction de variables aléatoires indépendantes) Un problème pratique important est de savoir si étant données deux variables aléatoires X et Y , on est capable de construire sur un même espace de probabilité deux autres variables X' et Y' indépendantes et de même lois que X et Y respectivement. La réponse est oui, et la solution est relativement simple. On utilise un espace produit. Supposons que X soit définie sur un espace (Ω_1, P_1) et Y sur un espace (Ω_2, P_2) . Alors on pose

$$\Omega = \Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1 \text{ et } \omega_2 \in \Omega_2\},$$

et on définit la probabilité P sur Ω par

$$P(\omega_1, \omega_2) = P_1(\omega_1) P_2(\omega_2).$$

(Vérifier que c'est bien une probabilité!) Puis on définit $Z : \Omega \rightarrow \mathbb{R}^2$, par

$$Z(\omega_1, \omega_2) = (X(\omega_1), Y(\omega_2)),$$

et on note π_1 et π_2 les projections de \mathbb{R}^2 dans \mathbb{R} :

$$\pi_1(x, y) = x \quad \text{et} \quad \pi_2(x, y) = y.$$

Alors on vérifie (exercice) que $X' = \pi_1 \circ Z$ et $Y' = \pi_2 \circ Z$ sont des variables sur (Ω, P) qui sont indépendantes et de même lois que X et Y respectivement.

Bien sûr cette construction peut s'étendre à un nombre fini quelconque de variables aléatoires (en fait même à un nombre dénombrable de variables, mais la preuve sort alors du cadre de ce cours).

Remarque (Retour sur la loi Binomiale) Soit $(X_i)_{1 \leq i \leq n}$ une suite de variables aléatoires indépendantes et de loi de Bernoulli de paramètre p , définies sur un même espace de probabilité (Ω, P) . Alors on peut noter (exercice!) que $S_n = X_1 + \dots + X_n$ suit une loi binomiale de paramètres n et p . Cette observation permet de retrouver aisément le résultat de la proposition 2.11 grâce à la proposition précédente. En effet on trouve directement

$$E[S_n] = E[X_1] + \dots + E[X_n] = np,$$

et

$$\text{Var}(S_n) = \text{Var}(X_1) + \dots + \text{Var}(X_n) = np(1 - p).$$

3.2 Probabilités Conditionnelles.

3.2.1 Motivation.

Nous allons introduire la notion de probabilité conditionnelle. Imaginons un sondage qui consiste à choisir un(e) français(e) au hasard, et à lui demander :

1. s'il(elle) aime le foot ;
2. s'il(elle) habite Marseille.

Si le sondage est fait sur n personnes, l'espace des issues est

$$\Omega = \{\omega = ((\omega_i, \tilde{\omega}_i), i \leq n), \omega_i \in \{F, \bar{F}\}, \tilde{\omega}_i \in \{M, \bar{M}\}\}.$$

$\omega_i = F$ signifie que la i -ème personne interrogée aime le foot ; $\tilde{\omega}_i = M$ signifie que la i -ème personne interrogée vit à Marseille.

Pour tout $\omega \in \Omega$,

$$P(\omega) = P_1(\omega_1, \tilde{\omega}_1) \dots P_1(\omega_n, \tilde{\omega}_n).$$

où P_1 est une probabilité sur $\Omega_1 = \{(F, M), (\bar{F}, M), (F, \bar{M}), (\bar{F}, \bar{M})\}$. $P_1((F, M))$ est la proportion de français qui vivent à Marseille et qui aiment le foot.

Intuitivement (cette intuition deviendra un théorème au chapitre 6), la proportion de gens dans **notre échantillon** qui vivent à Marseille et aiment le foot, devrait nous donner une idée de ce qu'est $P_1((F, M))$, du moins si notre échantillon est assez grand.

$$\frac{|\{i \leq n : \omega_i = F, \tilde{\omega}_i = M\}|}{n} \xrightarrow[n \rightarrow \infty]{} P_1(F, M).$$

Quelle est la proportion de Marseillais aimant le foot ? Intuitivement cela va correspondre à la proportion asymptotique d'amateurs de foot parmi les Marseillais de notre échantillon. En d'autres termes,

$$P(F|M) = \lim_{n \rightarrow \infty} \frac{|\{i \leq n : \omega_i = F, \tilde{\omega}_i = M, \}|}{|\{i \leq n : \tilde{\omega}_i = M\}|}$$

Mais,

$$\frac{|\{i \leq n : \omega_i = F, \tilde{\omega}_i = M\}|}{|\{i \leq n : \tilde{\omega}_i = M\}|} = \frac{|\{i \leq n : \omega_i = F, \tilde{\omega}_i = M\}|}{n} \frac{n}{|\{i \leq n : \tilde{\omega}_i = M\}|}.$$

Cette dernière quantité converge vers

$$\frac{P_1((F, M))}{P_1(\{\omega \in \Omega_1, \tilde{\omega}_1 = M\})}.$$

Cette discussion intuitive motive la définition suivante.

3.2.2 Probabilité conditionnelle.

Définition 3.9 Soit A et B deux évènements tels que $P(B) > 0$. La probabilité conditionnelle de A sachant B est

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

On dit aussi que $P(A|B)$ est la probabilité de A sachant B .

Proposition 3.10 $P(\cdot|B)$ est une probabilité sur Ω telle que $P(B|B) = 1$.

Preuve. Tout d'abord, notons que pour tout évènement $A \subset \Omega$, $P(A|B)$ est un nombre positif, en tant que rapport de deux nombres positifs. De plus, puisque $(A \cap B) \subset B$, $P(A \cap B) \leq P(B)$, et $P(A|B) \leq 1$.

Par ailleurs, $P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = 1$, et on a aussi $P(B|B) = 1$.

Enfin, si $(A_i)_{i \in I}$ est une famille d'évènements deux à deux disjoints, la famille $(A_i \cap B)_{i \in I}$ est encore une famille d'évènements deux à deux disjoints. Ainsi,

$$\begin{aligned} P(\cup_{i \in I} A_i | B) &= \frac{P((\cup_{i \in I} A_i) \cap B)}{P(B)} = \frac{P(\cup_{i \in I} (A_i \cap B))}{P(B)} \\ &= \frac{\sum_{i \in I} P(A_i \cap B)}{P(B)} = \sum_{i \in I} P(A_i | B). \end{aligned}$$

■

Proposition 3.11 Si A et B sont deux évènements indépendants, avec $P(B) > 0$,

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

Intuitivement, si deux évènements sont indépendants, savoir que B est réalisé ne nous apprend rien sur A .

Preuve.

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

■

Exemple 3.12 Supposons qu'on dispose de deux urnes contenant des boules blanches et noires. La première urne contient 100 boules, dont 99 blanches. La deuxième urne contient 100 boules, dont 10 blanches. Un jeu consiste à

1. choisir une urne au hasard ;
2. faire deux tirages successifs (avec remise) d'une boule dans l'urne choisie.

Si la première boule tirée est blanche, quelle est la probabilité que la deuxième soit également blanche ?

L'espace des issues est

$$\Omega = \{\omega = (n, \omega_1, \omega_2) : n \in \{1, 2\}, \omega_i \in \{B, N\}\}.$$

L'issue (n, ω_1, ω_2) signifie qu'on a choisi l'urne n , que la première boule tirée est de couleur ω_1 , et la seconde de couleur ω_2 . La probabilité d'une telle issue est

$$\begin{aligned} P(1, \omega_1, \omega_2) &= \frac{1}{2} P_1(\omega_1) P_1(\omega_2) \quad \text{où} \quad P_1(B) = \alpha = 0,99; \\ P(2, \omega_1, \omega_2) &= \frac{1}{2} P_2(\omega_1) P_2(\omega_2) \quad \text{où} \quad P_2(B) = \beta = 0,1. \end{aligned}$$

Notons par X_1 et X_2 les v.a. qui donnent respectivement la couleur de la première et de la deuxième boule tirée :

$$\begin{array}{lcl} X_1 : & \Omega & \rightarrow \{B, N\} \\ & \omega = (n, \omega_1, \omega_2) & \mapsto \omega_1 \end{array} \quad , \quad \begin{array}{lcl} X_2 : & \Omega & \rightarrow \{B, N\} \\ & \omega = (n, \omega_1, \omega_2) & \mapsto \omega_2. \end{array}$$

On a

$$\begin{aligned}
 P(X_2 = B|X_1 = B) &= \frac{P(\{X_1 = B\} \cap \{X_2 = B\})}{P(X_1 = B)} \\
 &= \frac{P(1, B, B) + P(2, B, B)}{P(1, B, B) + P(1, B, N) + P(2, B, B) + P(2, B, N)} \\
 &= \frac{\alpha^2/2 + \beta^2/2}{\alpha/2 + \beta/2} = \frac{\alpha^2 + \beta^2}{\alpha + \beta}.
 \end{aligned}$$

Par ailleurs, si on ne savait rien sur la première boule

$$P(X_2 = B) = P(\{(1, B, B), (1, N, B), (2, B, B), (2, N, B)\}) = \frac{1}{2}(\alpha + \beta).$$

On note que $P(X_2 = B|X_1 = B) > P(X_2 = B)$. Intuitivement, comme la première boule tirée est blanche, l'urne choisie a plus de chance d'être l'urne 1 que l'urne 2.

3.2.3 Formules des probabilités totales et des probabilités composées.

Proposition 3.13 *Formule des probabilités totales, ou de Bayes.*

Soit $\{A_i\}_{i \in I}$ une partition de Ω , telle que pour tout i , $P(A_i) > 0$. Pour tout évènement $B \subset \Omega$,

$$P(B) = \sum_{i \in I} P(B|A_i)P(A_i).$$

Preuve. Notons que $B = B \cap (\cup_{i \in I} A_i) = \cup_{i \in I} (B \cap A_i)$. Comme les A_i sont disjoints, les $A_i \cap B$ le sont aussi. Donc,

$$P(B) = \sum_{i \in I} P(B \cap A_i) = \sum_{i \in I} P(B|A_i)P(A_i).$$

■

Proposition 3.14 *Formule des probabilités composées.*

Soient A_1, \dots, A_n des évènements tels que $P(A_1 \cap A_2 \cap \dots \cap A_{n-1}) > 0$.

$$P(A_1 \cap \dots \cap A_n) = P(A_n|A_1 \cap \dots \cap A_{n-1})P(A_{n-1}|A_1 \cap \dots \cap A_{n-2}) \dots P(A_2|A_1)P(A_1).$$

Preuve. La démonstration peut se faire par récurrence sur n . Pour $n = 2$, il s'agit simplement de la définition de la probabilité conditionnelle. Supposons la formule vraie à l'ordre n et montrons la à l'ordre $n + 1$.

$$\begin{aligned}
 P(A_1 \cap \dots \cap A_{n+1}) &= P(A_{n+1}|A_1 \cap \dots \cap A_n)P(A_1 \cap \dots \cap A_n) \text{ par définition,} \\
 &= P(A_{n+1}|A_1 \cap \dots \cap A_n)P(A_n|A_1 \cap \dots \cap A_{n-1}) \dots P(A_2|A_1)P(A_1),
 \end{aligned}$$

par hypothèse de récurrence.

■

Chapitre 4

Fonction génératrice.

4.1 Fonction génératrice.

Définition 4.1 Soit X une variable aléatoire prenant ses valeurs dans \mathbb{N} . La **fonction génératrice** de X est la fonction continue G_X définie pour tout $t \in [-1, 1]$ par :

$$G_X(t) = E[t^X] = \sum_{k=0}^{\infty} t^k P(X = k).$$

Remarque La fonction G_X est bien définie pour tout $t \in [-1, +1]$, puisque $|t^k P[X = k]| \leq P[X = k]$, et $\sum_{k=0}^{\infty} P[X = k] = 1$. Par ailleurs, dans la définition, la deuxième égalité découle, pour $t \neq 0$, de la formule $E[f(X)] = \sum_k f(k)P(X = k)$, valable pour toute fonction f (telle que l'espérance de $f(X)$ soit bien définie). Pour $t = 0$, on peut prendre l'égalité comme une convention, qui rend la fonction G_X continue en 0.

On observera par ailleurs, que si la variable X ne prend qu'un nombre fini de valeurs, alors G_X est définie pour tout $t \in \mathbb{R}$, et est un polynôme.

On notera les formules importantes :

$$G_X(1) = 1, \text{ et } G_X(0) = P(X = 0).$$

Maintenant en dérivant sous le signe "somme", on obtient

$$G'_X(t) = \sum_{k=1}^{\infty} kt^{k-1} P(X = k).$$

Cette dernière identité est claire si X prend un nombre fini de valeurs. Elle reste vraie dans le cas général pour $t \in]-1; +1[$. Ainsi,

$$G'_X(0) = P(X = 1).$$

De la même façon, en dérivant k fois, on obtient

$$\forall k \in \mathbb{N} \quad P(X = k) = \frac{G_X^{(k)}(0)}{k!}.$$

Cette expression montre qu'il suffit de connaître G_X au voisinage de 0 pour caractériser la loi de X . En particulier on a le résultat suivant :

Proposition 4.2 *La fonction génératrice d'une variable aléatoire X à valeurs dans \mathbb{N} caractérise sa loi. Autrement dit, si X et Y sont deux variables à valeurs dans \mathbb{N} ,*

$$G_X = G_Y \Leftrightarrow \forall k \in \mathbb{N}, P(X = k) = P(Y = k).$$

On notera maintenant que si X prend un nombre fini de valeurs, en prenant $t = 1$ dans les expressions de G'_X et G''_X , on obtient,

$$E[X] = G'_X(1) \quad \text{et} \quad \text{Var}(X) = G''_X(1) + G'_X(1) - G'_X(1)^2.$$

Plus généralement, si X prend un nombre dénombrable de valeurs, et si son espérance et sa variance sont bien définies, alors :

$$E[X] = \lim_{t \rightarrow 1, t < 1} G'_X(t) \quad \text{et} \quad \text{Var}(X) = \lim_{t \rightarrow 1, t < 1} (G''_X(t) + G'_X(t) - G'_X(t)^2).$$

Le résultat suivant est très utile :

Proposition 4.3 *Soient X et Y deux variables aléatoires **indépendantes**. Alors*

$$G_{X+Y} = G_X G_Y.$$

Preuve. La proposition 3.7 nous dit que pour tout $t \neq 0$ fixé, les variables t^X et t^Y sont indépendantes. On déduit alors de la proposition 3.8 que $G_{X+Y}(t) = G_X(t) G_Y(t)$, pour tout $t \neq 0$. L'égalité en $t = 0$ s'en déduit par continuité. ■

Exemple 4.4 Si X est une variable de **Bernoulli** de paramètre p ,

$$G_X(t) = t^0 P(X = 0) + t^1 P(X = 1) = (1 - p) + pt.$$

Exemple 4.5 Si S_n est une variable de loi **binomiale** ($S_n \sim B(n, p)$), (pensez-y comme à la somme de n variables indépendantes de Bernoulli X_1, \dots, X_n), alors en utilisant l'indépendance des X_i et la proposition 4.3,

$$G_{S_n}(t) = G_{X_1}(t)^n = ((1 - p) + pt)^n.$$

En utilisant la formule du binôme, on obtient

$$G_{S_n}(t) = \sum_{k=1}^n C_n^k (1 - p)^{n-k} (pt)^k.$$

En regardant le coefficient de t^k , on retrouve que

$$P(S_n = k) = C_n^k p^k (1 - p)^{n-k}.$$

Exemple 4.6 Si $X \sim \mathcal{P}(\lambda)$ est une variable de **Poisson**, alors

$$G_X(t) = \sum_{k=0}^{\infty} t^k P(X = k) = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} t^k = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(t\lambda)^k}{k!} = e^{\lambda(t-1)}.$$

Exemple 4.7 Soient X et Y deux variables indépendantes de Poisson de paramètres λ et μ , respectivement. Quelle est la distribution de $X + Y$? Calculons la fonction génératrice de $X + Y$.

$$G_{X+Y}(t) = G_X(t)G_Y(t) \text{ par indépendance de } X \text{ et } Y.$$

Par le calcul précédent, on a donc

$$G_{X+Y}(t) = e^{\lambda(t-1)} e^{\mu(t-1)} = e^{(\lambda+\mu)(t-1)}.$$

On reconnaît la fonction génératrice d'une variable de Poisson de paramètre $\lambda + \mu$ et la proposition 4.2 nous permet de conclure que $X + Y$ est une variable de Poisson de paramètre $\lambda + \mu$.

4.2 Somme d'un nombre aléatoire de variables aléatoires.

Une poule pond N oeufs. On suppose que N est une variable de Poisson $\mathcal{P}(\lambda)$. Chaque oeuf, indépendamment de tout le reste, éclot avec probabilité p . Soit K le nombre de poussins. Quelle est la loi de K ?

Pour répondre de façon rigoureuse à cette question, il faut d'abord définir Ω . Une issue nous informera sur le nombre d'oeufs pondus, et l'état de chaque oeuf. Intuitivement, on voudrait écrire $\omega = (n, o_1, \dots, o_n)$ où n est un entier correspondant au nombre d'oeufs pondus et

$$o_i = \begin{cases} 1 & \text{si le } i\text{-ème oeuf donne un poussin;} \\ 0 & \text{sinon.} \end{cases}$$

En pratique, il est malaisé de traiter avec des issues dont la taille est variable : $(1, 0)$ et $(3, 1, 0, 1)$... En première lecture, vous pouvez directement aller à la définition de la probabilité de pondre n_0 oeufs de type $(o_1, o_2, \dots, o_{n_0})$, qui est donnée dans l'équation (4.1). Une façon (artificielle) très commode est d'introduire l'espace des états d'une suite infinie d'oeufs pondus

$$\tilde{\Omega} = \{\tilde{\omega} = \{\tilde{\omega}_i, i \in \mathbb{N}\}, \tilde{\omega}_i \in \{0, 1\}\},$$

C'est un peu comme si on se plaçait dans la situation où notre poule pouvait a priori pondre un nombre infini d'oeufs! Puis, on introduit une probabilité \tilde{P} sur $\tilde{\Omega}$, telle que pour chaque entier k et chaque choix $o_1, \dots, o_k \in \{0, 1\}$

$$\tilde{P}(\{\tilde{\omega} : \tilde{\omega}_1 = o_1, \dots, \tilde{\omega}_k = o_k\}) = p^{S_k(o)}(1-p)^{k-S_k(o)},$$

où $S_k(o) = o_1 + \dots + o_k$ est le nombre de poussins parmi o_1, o_2, \dots, o_k . Ceci correspondrait à la probabilité d'avoir (o_1, o_2, \dots, o_k) sachant que notre poule a pondu k oeufs. Finalement, on écrit notre vrai espace en tenant compte du nombre d'oeufs pondus

$$\Omega = \{\omega = (n, \tilde{\omega}), n \in \mathbb{N}, \tilde{\omega} \in \tilde{\Omega}\}.$$

La probabilité associée à l'issue (n_0, ω^o) est

$$P((n_0, \omega^o)) = e^{-\lambda} \frac{\lambda^{n_0}}{n_0!} \tilde{P}(\{\tilde{\omega} : \tilde{\omega}_1 = \omega_1^o, \dots, \tilde{\omega}_{n_0} = \omega_{n_0}^o\}). \quad (4.1)$$

On définit sur l'espace Ω , la variable de Poisson $N(n, \tilde{\omega}) = n$ et les variables de Bernoulli

$$\forall i \in \mathbb{N}, \quad X_i(n, \tilde{\omega}) = \tilde{\omega}_i.$$

(Vérifier que N et les $\{X_n, n \in \mathbb{N}\}$ sont indépendants.)

On peut alors définir la variable K représentant le nombre de poussins obtenus, par

$$K(\omega) = \begin{cases} 0 & \text{si } N(\omega) = 0, \\ X_1(\omega) + \cdots + X_n(\omega) & \text{si } N(\omega) = n \end{cases}$$

c'est-à-dire $K(\omega) = X_1(\omega) + \cdots + X_{N(\omega)}(\omega)$. La difficulté de cette expression est que l'on somme un nombre aléatoire de X_i . Il est plus simple de l'exprimer comme

$$K = \sum_{k=1}^{\infty} \mathbb{1}_{\{N=k\}} (X_1 + \cdots + X_k).$$

Calculons $G_K(t)$. Notons d'abord que

$$t^K = \mathbb{1}_{\{N=0\}} + \sum_{n=1}^{\infty} \mathbb{1}_{\{N=n\}} t^{(X_1 + \cdots + X_n)}.$$

Ainsi, en utilisant que l'espérance d'une somme de termes est la somme des espérances de ces termes (ce qui marche même pour une somme infinie de termes positifs)

$$\begin{aligned} G_K(t) &= E[t^K] = E[\mathbb{1}_{\{N=0\}}] + E\left[\sum_{n=1}^{\infty} \mathbb{1}_{\{N=n\}} t^{X_1 + \cdots + X_n}\right] \\ &= P(N=0) + \sum_{n=1}^{\infty} E\left[\mathbb{1}_{\{N=n\}} t^{X_1 + \cdots + X_n}\right] \\ &= P(N=0) + \sum_{n=1}^{\infty} P(N=n) (E[t^{X_1}])^n \\ &= \sum_{n=0}^{\infty} P(N=n) (G_{X_1}(t))^n \\ &= G_N(G_{X_1}(t)). \end{aligned} \tag{4.2}$$

Après ce long calcul, utilisons les expressions de G_X et G_N que nous avons obtenues respectivement dans l'exemple 4.4 et 4.6.

$$G_{X_1}(t) = 1 - p + pt, \quad \text{et} \quad G_N(y) = \exp(\lambda(y-1)).$$

On obtient

$$G_K(t) = G_N(G_{X_1}(t)) = \exp(\lambda(1 - p + pt - 1)) = \exp(p\lambda(t-1)).$$

On reconnaît la fonction génératrice d'une loi de Poisson de paramètre $p\lambda$. Par la proposition 4.2, on en conclut que K est une variable de Poisson de paramètre $p\lambda$.

Maintenant, le nombre moyen de poussins s'obtient en dérivant une fois $G_K(t)$ par rapport à t

$$E[K] = G'_K(1) = \lambda p = E(N)p.$$

Cette dernière égalité est assez intuitive : le nombre moyen de poussins obtenus est égal au nombre moyen d'oeufs pondus, multiplié par la probabilité qu'à chaque oeuf d'éclore.

De la même façon, on montre la proposition suivante.

Proposition 4.8 Soient $\{X_n, n \geq 1\}$ une suite de variables indépendantes de même loi, et N une variable à valeurs entières indépendante des $\{X_n, n \geq 1\}$, alors

$$S = \begin{cases} 0, & \text{si } N = 0 \\ X_1 + \cdots + X_N, & \text{si } N > 0 \end{cases}$$

a pour fonction génératrice $G_S(t) = G_N(G_X(t))$, et pour espérance $E(S) = E(N)E(X)$.

Preuve. Le résultat concernant la fonction génératrice de S se démontre exactement de la même façon que dans l'exemple des oeufs. Pour obtenir l'espérance, on écrit

$$E(S) = G'_S(1) = G'_N(G_X(1))G'_X(1) = G'_N(1)G'_X(1) = E(N)E(X).$$

■

4.3 Evolution des Populations

4.3.1 Motivation

Nous allons décrire un modèle d'évolution d'une population, et montrer en quoi les fonctions génératrices sont utiles. La génération 0 est composée d'un individu. Cet individu fait $\eta_1^{(1)}$ enfants ; $\eta_1^{(1)}$ est une variable aléatoire, dont la loi est appelée *loi de la progéniture*

$$P(\eta_1^{(1)} = k) = p_k, \quad \forall k \in \mathbb{N}.$$

Ces enfants forment la génération 1. Chacun d'eux, indépendamment du reste, fait un nombre aléatoire d'enfants : $\eta_1^{(2)}$ est le nombre d'enfants du 1er, $\eta_2^{(2)}$ est le nombre d'enfants du 2ème, etc...L'indice supérieur 2 rappelle que ces enfants forment la génération 2. Et ainsi de suite...

Soit Z_n le nombre d'individus à la génération n . Pour obtenir Z_{n+1} , il faut additionner le nombre d'enfants de chaque individu de la génération n : si $Z_n = 0$ alors $Z_{n+1} = 0$, sinon

$$Z_{n+1} = \eta_1^{(n+1)} + \eta_2^{(n+1)} + \dots + \eta_{Z_n}^{(n+1)}.$$

La façon la plus simple de comprendre cette somme avec un indice aléatoire est de la *décomposer* sur toutes les valeurs possibles de Z_n

$$Z_{n+1} = 1_{\{Z_n=1\}}(\eta_1^{(n+1)}) + \dots + 1_{\{Z_n=k\}}(\eta_1^{(n+1)} + \dots + \eta_k^{(n+1)}) + \dots$$

Quelle est la probabilité que la population finisse par s'éteindre ?

4.3.2 Probabilité d'extinction.

Proposition 4.9 *Soit la suite de variables entières $\{Z_n, n \geq 0\}$ définie plus haut. Alors la probabilité que la population finisse par s'éteindre est le plus petit réel positif, disons $\chi > 0$, tel que $G_{Z_1}(\chi) = \chi$.*

Preuve. Nous exprimons l'évènement $A :=$ "la population finit par s'éteindre" en terme des $\{Z_n\}$.

$$A = \cup_{n=1}^{\infty} \{Z_n = 0\}.$$

Comme $\{Z_n = 0\} \subset \{Z_{n+1} = 0\}$, on peut appliquer la propriété 5/ de la proposition 1.7

$$P(A) = \lim_{n \rightarrow \infty} P(\{Z_n = 0\}) = \lim_{n \rightarrow \infty} G_{Z_n}(0) \leq 1.$$

Cette limite existe bien et nous l'appelons χ . Nous avons donc besoin de calculer $G_{Z_n}(0)$. Pour $n > 1$, on utilise le fait que Z_n et les $\{\eta_k^{(n+1)}, k = 1, \dots\}$ sont indépendants.

$$G_{Z_{n+1}}(z) = \sum_{k=1}^{\infty} E[1_{\{Z_n=k\}} z^{\eta_1^{(n+1)} + \dots + \eta_k^{(n+1)}}] = \sum_{k=1}^{\infty} P(Z_n = k) (E[z^{\eta_1^{(1)}}])^k = G_{Z_n}(G_{Z_1}(z)).$$

Donc, si on pose $f(z) := G_{Z_1}(z)$, $g_1(z) = f(z)$ et $g_n(z) := G_{Z_n}(z)$, alors

$$g_{n+1}(z) = g_n(f(z)) \implies g_n(z) = f^{(n)}(z) := f(f(f(\dots n \text{ fois } \dots f(z) \dots))).$$

Notons que f est continue sur $[0, 1]$, et donc

$$\lim_{n \rightarrow \infty} g_n(0) = \chi \implies \lim_{n \rightarrow \infty} f(g_n(0)) = f(\chi).$$

Or, comme $f(g_n) = g_{n+1}$, on a que $f(\chi) = \chi$. Il nous reste à voir que χ est le plus petit réel positif vérifiant $f(\chi) = \chi$. Soit $\zeta \in [0, 1]$ tel que $f(\zeta) = \zeta$. Montrons que $\chi \leq \zeta$. Pour cela, il suffit de remarquer que f est croissante sur $[0, 1]$. En effet, si $0 \leq z_1 \leq z_2 \leq 1$, alors $0 \leq z_1^{Z_1} \leq z_2^{Z_1} \leq 1$, et en passant aux espérances, $f(z_1) \leq f(z_2)$. Par conséquent, si $\zeta \in [0, 1]$ est tel que $f(\zeta) = \zeta$, $\zeta = f(\zeta) \geq f(0)$, et en prenant f de chaque coté $\zeta \geq f^{(2)}(0)$. En répétant le procédé, on a que $\zeta \geq f^{(n)}(0)$ et donc en passant à la limite $\zeta \geq \chi$. ■

4.3.3 Exemple.

Supposons que la loi de la progéniture soit définie par

$$P(\eta_1^{(1)} = k) = P(Z_1 = k) = p(1-p)^k, \quad \text{pour } k = 0, 1, \dots,$$

Donc

$$G_{Z_1}(z) = \sum_{n=0}^{\infty} z^n P(\eta_1^{(1)} = n) = \frac{p}{1 - z(1-p)}.$$

En particulier $G_{Z_1}(0) = p$ et $G'_{Z_1}(1) = (1-p)/p$. On cherche la plus petite solution de $G_{Z_1}(\chi) = \chi$. On résoud

$$\frac{p}{1 - z(1-p)} = z \implies 0 = z^2(1-p) - z + p = (1-p)(z-1)\left(z - \frac{p}{1-p}\right).$$

Donc si $p/(1-p) < 1$ on a $\chi = p/(1-p)$, sinon $\chi = 1$. Ainsi, la population finit par s'éteindre avec probabilité 1 lorsque $p \geq 1/2$, c'est-à-dire lorsque le nombre moyen d'enfants est plus petit ou égal à 1.

Exercice : Démontrer en général que si $E[Z_1] > 1$, alors $\chi < 1$, et que si $E[Z_1] < 1$, alors $\chi = 1$. Dans le cas où $E[Z_1] = 1$, et où $P(Z_1 = 0) \neq 0$, démontrer également que $\chi = 1$. (On pourra utiliser le fait que G'_X est croissante sur $[0, 1]$.)

Chapitre 5

Variables continues. Loi normale.

5.1 Axiomatique des probabilités.

Jusqu'à présent, nous n'avons considéré que des phénomènes aléatoires prenant un nombre fini ou dénombrable de valeurs. Dans bien des situations, ce n'est pas le cas. Par exemple, considérons une particule qui tombe sur une plaque carrée de longueur L . L'espace des issues est ici $\Omega = [0, L]^2$. Dans ce cas, il est impossible de construire une probabilité P sur Ω , comme nous l'avons fait jusqu'à présent en associant une probabilité à chaque issue possible. Ainsi, dans l'exemple précédent, et si la probabilité de tomber dans un sous-ensemble A de Ω est proportionnelle à l'aire de A , alors $P(\omega) = 0$ pour tout $\omega \in \Omega$. On doit vraiment voir ici une probabilité comme une fonction sur une famille \mathcal{A} de sous-ensembles de Ω (dans l'exemple précédent, si $A \subset \Omega$, $P(A) = \frac{\text{aire de } A}{L^2}$). Puisqu'on veut pouvoir calculer la probabilité de la réunion d'évènements, la famille \mathcal{A} considérée doit être stable par réunion. On arrive ainsi à la définition suivante.

Définition 5.1 Soit Ω un espace d'issues. On appelle **tribu** d'évènements, toute famille \mathcal{A} de sous-ensembles de Ω vérifiant

1. $\Omega \in \mathcal{A}$.
2. Si $A \in \mathcal{A}$, $A^c \in \mathcal{A}$.
3. Pour tout famille finie ou dénombrable $(A_i)_{i \in I}$ d'éléments de \mathcal{A} , $\cup_{i \in I} A_i \in \mathcal{A}$.

Les propriétés 2 et 3 entraînent que toute intersection dénombrable d'éléments de \mathcal{A} est encore dans \mathcal{A} .

Exemple 5.2 1. L'ensemble des parties de Ω est une tribu. C'est la plus grande. C'est celle que nous avons de façon implicite toujours considéré dans le cas où Ω est fini ou dénombrable.

2. $\{\emptyset, \Omega\}$ est une tribu. C'est la plus petite.
3. Si $A \subset \Omega$, $\{\emptyset, A, A^c, \Omega\}$ est une tribu. C'est la plus petite tribu contenant A . On dit que c'est la tribu engendrée par A .

Définition 5.3 Soit Ω un espace d'issues, et \mathcal{A} une tribu sur Ω . Une **probabilité** sur (Ω, \mathcal{A}) est une application $P : \mathcal{A} \rightarrow [0, 1]$ telle que

1. $P(\Omega) = 1$;

2. Pour toute famille finie ou dénombrable $(A_i)_{i \in I}$ d'éléments de \mathcal{A} deux à deux disjoints,

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

Le triplet (Ω, \mathcal{A}, P) est appelé un **espace de probabilité**.

Dans le cas fini ou dénombrable, on prendra toujours $\mathcal{A} = \mathcal{P}(\Omega)$ et on retrouve la notion de probabilité du chapitre 1.

On peut vérifier que toutes les propriétés d'une probabilité vues dans le cas fini ou dénombrable restent vraies dans le cadre d'un espace de probabilité général. On peut aussi redéfinir dans ce cadre la notion d'évènements indépendants, de probabilité conditionnelle, etc...

5.2 Variables aléatoires.

5.2.1 Définition.

Définition 5.4 Une **variable aléatoire** (v.a.) sur (Ω, \mathcal{A}) est une application $X : \Omega \rightarrow \mathbb{R}$ telle que pour tout $x \in \mathbb{R}$, $\{X \leq x\} \in \mathcal{A}$.

Remarque Si X est une variable aléatoire, on peut parler de $P(X \leq x)$. La fonction $x \in \mathbb{R} \mapsto F_X(x) = P(X \leq x)$ est la **fonction de répartition** de X . On montre exactement comme dans le cas discret que les résultats de la proposition 1.17 restent vrais dans le cadre général. En particulier F_X est croissante de \mathbb{R} dans $[0, 1]$, continue à droite, tend vers 0 en $-\infty$ et vers 1 en $+\infty$.

Notons que $\{y < X \leq x\} = \{X \leq x\} \cap \{X \leq y\}^c$ est un élément de la tribu \mathcal{A} . On peut donc parler de $P(y < X \leq x)$ et par les axiomes des probabilités, on a $P(y < X \leq x) = F_X(x) - F_X(y)$.

De même, $\{y \leq X \leq x\} = \bigcap_n \{y - 1/n < X \leq x\}$ est dans la tribu \mathcal{A} . De plus, comme les évènements $\{y - 1/n < X \leq x\}$ sont décroissants, on a

$$P(y \leq X \leq x) = \lim_{n \rightarrow \infty} P(y - 1/n < X \leq x) = \lim_{n \rightarrow \infty} F_X(x) - F_X(y - 1/n).$$

De la même façon, on peut vérifier que pour tout intervalle I (ouvert, fermé, semi-ouvert) de \mathbb{R} , $\{X \in I\}$ est un élément de \mathcal{A} (et il en est donc de même de toute réunion dénombrable de tels intervalles).

Définition 5.5 Soit X une variable aléatoire définie sur (Ω, \mathcal{A}, P) . On appelle **loi de probabilité de X** la fonction qui à tout intervalle I associe le nombre $P(X \in I)$.

Comme on vient de le voir, la loi d'une v.a. s'exprime en fonction de la fonction de répartition. Ainsi,

Proposition 5.6 La fonction de répartition de X détermine la loi de X .

5.2.2 Variables à densité.

Définition 5.7 On dit que X est une variable aléatoire à densité s'il existe une fonction f positive continue par morceaux, telle que $\forall x \in \mathbb{R}$,

$$P(X \leq x) = \int_{-\infty}^x f(t) dt$$

On dit alors que X est de densité f , ou inversement que f est la densité de X .

Proposition 5.8 Si X est de densité f ,

1. pour tout $x \in \mathbb{R}$, $P[X = x] = 0$;
2. pour tous $x, y \in \mathbb{R}$ tels que $x \leq y$, $P[X \in (x; y)] = \int_x^y f(t) dt$; en particulier $\int_{-\infty}^{+\infty} f(t) dt = 1$.
3. La fonction de répartition F_X de X est dérivable et $F'_X = f$ (sauf éventuellement là où f n'est pas continue).

Preuve. $\{X = x\} = \bigcap_n \{x - 1/n < X \leq x\}$, et l'intersection est décroissante. On a donc

$$\begin{aligned} P[X = x] &= \lim_{n \rightarrow \infty} P[x - 1/n < X \leq x] = \lim_{n \rightarrow \infty} F_X(x) - F_X(x - 1/n) \\ &= \lim_{n \rightarrow \infty} \int_{x-1/n}^x f(t) dt = 0, \quad \text{par les propriétés de l'intégrale.} \end{aligned}$$

Ainsi, quel que soit le type d'intervalle considéré (ouvert, semi-ouvert, fermé), on a

$$P[X \in (x; y)] = P[X \in]x; y] = F_X(y) - F_X(x) = \int_x^y f(t) dt.$$

On a donc pour $\epsilon > 0$ petit, $F(x + \epsilon) - F(x) = P[X \in (x, x + \epsilon)] = \int_x^{x+\epsilon} f(t) dt \simeq f(x)\epsilon$, et le point 3. en découle. Enfin $\int_{-\infty}^{+\infty} f(t) dt = P[-\infty < X < \infty] = 1$. ■

Remarque Une fonction positive f continue par morceaux telle que $\int_{-\infty}^{+\infty} f(t) dt = 1$, est appelée **densité de probabilité**. On peut montrer que toute densité de probabilité est la densité d'une variable aléatoire.

Remarque Puisque $P[X \in (x - \epsilon; x + \epsilon)] \simeq 2f(x)\epsilon$, on peut penser à $f(x)$ comme à une mesure de la probabilité pour X de valoir x , même si $P(X = x) = 0$. Mais attention toutefois au fait que l'on peut avoir $f(x) > 1$, pour certains x ...

Exemple 5.9 (Loi uniforme) *licences.st@univ-provence.fr* Soit $a < b$ deux réels et soit f la fonction valant $1/(b - a)$ sur $[a, b]$, et 0 partout ailleurs. f est positive, continue (sauf en a et b) et $\int f(t) dt = 1$. f est donc une densité de probabilité. Une variable X de densité f est appelée **variable uniforme sur** $[a, b]$, et on note $X \sim \mathcal{U}([a, b])$. Par exemple si $X \sim \mathcal{U}([0, 1])$, la probabilité pour que $X \in [1/4, 1/2]$ vaut

$$P[X \in [1/4, 1/2]] = \int_{1/4}^{1/2} dt = 1/2 - 1/4 = 1/4.$$

Exemple 5.10 (Loi exponentielle) Soit λ un réel strictement positif et soit $f(x) = \lambda \exp(-\lambda x)$ pour $x \geq 0$ et $f(x) = 0$ pour $x < 0$. On vérifie que f est une densité de probabilité. En effet, f est positive, et

$$\int_{-\infty}^{\infty} f(x) dx = \int_0^{\infty} \lambda \exp(-\lambda x) dx = [-\exp(-\lambda x)]_0^{\infty} = 1.$$

Une variable aléatoire de densité f est une **variable exponentielle de paramètre λ** , et on note $X \sim \mathcal{E}(\lambda)$. On peut, par exemple, calculer la probabilité que $X > s$, pour $s > 0$

$$P(X > s) = \int_s^{\infty} \lambda \exp(-\lambda x) dx = \exp(-\lambda s).$$

Maintenant sachant que $X > s$, quelle est la probabilité que $X > t + s$, pour t et s positifs? Il s'agit de faire un calcul d'espérance conditionnelle.

$$\begin{aligned} P(X > t + s | X > s) &= \frac{P(\{X > t + s\} \cap \{X > s\})}{P(X > s)} = \frac{P(X > t + s)}{P(X > s)} \\ &= \frac{\exp(-\lambda(t + s))}{\exp(-\lambda s)} = \exp(-\lambda t) = P(X > t). \end{aligned}$$

On dit qu'une variable exponentielle est **sans mémoire**. C'est l'analogie continue de la loi géométrique, qui possède aussi cette propriété. Par exemple si X représente la durée d'une conversation téléphonique, la propriété précédente dit que lorsque la conversation a duré s minutes, la probabilité qu'elle dure t minutes de plus est la même que si la conversation venait de commencer. Ainsi une variable exponentielle ne permet de modéliser une conversation téléphonique qu'entre deux personnes très bavardes, et peu soucieuses de la tarification.

En pratique on utilise plutôt les v.a. exponentielles pour modéliser la durée de vie d'appareils ménagers ou d'ampoules électriques, ou encore le temps d'attente entre deux clients à un guichet.

Exemple 5.11 (Changement de variables) Si f est la densité de X , quelle est la densité de $2X + 1$? Posons $Y = 2X + 1$, et cherchons une fonction positive g telle que pour tous $a < b$

$$P[a < Y < b] = \int_a^b g(x) dx.$$

Il suffit de réécrire Y en fonction de X :

$$P[a < Y < b] = P[a < 2X + 1 < b] = P[(a - 1)/2 < X < (b - 1)/2] = \int_{(a-1)/2}^{(b-1)/2} f(x) dx.$$

Pour terminer, il faut transformer cette dernière intégrale, de façon que les bornes d'intégration soient a et b . On fait donc le changement de variable $y = 2x + 1$ (ou encore $x = (y - 1)/2$) de façon à ce que quand $x = (a - 1)/2$ alors $y = a$, et lorsque $x = (b - 1)/2$ alors $y = b$. Ainsi

$$\int_{(a-1)/2}^{(b-1)/2} f(x) dx = \int_a^b f\left(\frac{y-1}{2}\right) \frac{1}{2} dy.$$

Et $g(y) = f((y - 1)/2)/2$.

Exercice : Plus généralement montrer que si f est la densité de X , alors pour tout $\sigma, \mu \in \mathbb{R}$, $Y = \sigma X + \mu$ a pour densité

$$f_Y(t) = \frac{1}{\sigma} f\left(\frac{t - \mu}{\sigma}\right).$$

En particulier, si $X \sim \mathcal{U}([0, 1])$, montrer que $Y \sim \mathcal{U}([\mu, \mu + \sigma])$, et si $X \sim \mathcal{E}(\lambda)$, montrer que $\sigma X \sim \mathcal{E}(\lambda/\sigma)$.

5.2.3 Indépendance.

Définition 5.12 Deux variables X et Y définies sur un espace de probabilité (Ω, \mathcal{A}, P) sont **indépendantes** si pour tous intervalles I et J ,

$$P(\{X \in I\} \cap \{Y \in J\}) = P(X \in I)P(Y \in J).$$

Comme dans le cas discret, on a

Proposition 5.13 (admise) Si X et Y sont deux variables indépendantes, et si f et g sont deux fonctions continues, les variables $f(X)$ et $g(Y)$ sont indépendantes.

5.2.4 Espérance.

Nous avons vu dans le cas discret que si $\{x_k; k = 1, \dots\}$ est l'ensemble des valeurs prises par une variable X , alors pour toute fonction h

$$E[h(X)] = \sum_k h(x_k)P[X = x_k],$$

dès que la série est bien définie. Par ailleurs, si X est une variable de densité continue f , pour une quantité dx infinitésimale,

$$“f(x)dx = \int_x^{x+dx} f(u)du = P(X \in]x, x + dx[)”.$$

Par analogie avec le cas discret on voudrait écrire, pour une fonction h quelconque

$$“E[h(X)] = \sum_x h(x)P(X \in [x, x + dx]) = \sum_x h(x)f(x)dx”.$$
 (5.1)

Le sens rigoureux de (5.1) est donné par l'intégrale :

$$E[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)dx,$$
 (5.2)

dès que cette intégrale est bien définie.

Définition 5.14 Soit X une variable de densité f , et h une fonction continue. On définit l'**espérance de $h(X)$** par

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

dès que cette intégrale est bien définie.

Ainsi, l'espérance de X est

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

et la variance de X est

$$\text{Var}(X) = E[X^2] - (E[X])^2 = \int_{-\infty}^{\infty} x^2f(x)dx - \left(\int_{-\infty}^{\infty} xf(x)dx \right)^2.$$

Les propriétés de l'espérance que nous avons montrées dans le cas discret, restent vraies pour des variables continues. Nous en énonçons quelques-unes sans démonstration.

Proposition 5.15 (admise) Soit X et Y deux variables aléatoires et $\lambda \in \mathbb{R}$. Alors,

1. $E[\lambda X] = \lambda E[X]$.
2. Si $X \geq 0$, $E(X) \geq 0$.
3. $E[X + Y] = E[X] + E[Y]$.
4. Si X et Y sont indépendantes, alors $E[XY] = E[X]E[Y]$.
5. Si X et Y sont indépendantes, alors $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$.

Exemples : (à faire en exercice!)

$$\begin{aligned} \text{Si } X \sim \mathcal{U}([a, b]), \text{ alors } E(X) &= \frac{a+b}{2}, \text{ Var}(X) = \frac{(b-a)^2}{12} \\ \text{si } X \sim \mathcal{E}(\lambda), \text{ alors } E(X) &= \frac{1}{\lambda}, \text{ Var}(X) = \frac{1}{\lambda^2}. \end{aligned}$$

5.3 Lois Normales

Définition 5.16 La loi normale standard (ou loi normale centrée réduite) est la loi de densité

$$f(x) = \exp(-x^2/2)/\sqrt{2\pi}. \tag{5.3}$$

Une variable aléatoire de loi normale standard est aussi appelée variable normale ou variable gaussienne centrée réduite. Si X est une telle variable, on note $X \sim \mathcal{N}(0, 1)$.

Soit $a < b$ des nombres réels. On a donc

$$P(a < X < b) = \int_a^b f(x) dx = \int_a^b \exp(-x^2/2) \frac{dx}{\sqrt{2\pi}}.$$

Comme on n'a pas d'expression explicite pour la primitive de $\exp(-x^2/2)$, ces probabilités ne peuvent être calculées de façon exacte. On trouve des valeurs approchées dans les tables de la loi normale. En revanche on peut montrer que $\int_{-\infty}^{\infty} \exp(-x^2/2) dx = \sqrt{2\pi}$, et donc que l'on a bien $\int f(x) dx = 1$ (autrement dit f est bien une densité de probabilité).

Proposition 5.17 Soit X une variable aléatoire de loi normale standard, et F sa fonction de répartition. Alors

1. $E[X] = 0$.

2. $\text{Var}(X) = 1$.
3. $\forall x, P(X \leq -x) = P(X \geq x)$.
4. $\forall x \geq 0, P(-x \leq X \leq x) = 2F(x) - 1$.

Preuve. Notons f la densité de la loi normale.

1.

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx.$$

Comme f est une fonction paire ($f(x) = f(-x)$), $x \mapsto xf(x)$ est impaire. Donc $E(X) = 0$.

2.

$$\text{Var}(X) = E[X^2] = \int_{-\infty}^{\infty} x^2 \exp(-x^2/2) \frac{dx}{\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u(x)v'(x) dx,$$

en posant $u(x) = x$, $v(x) = -\exp(-x^2/2)$. Par intégration par parties,

$$\text{Var}(X) = \frac{1}{\sqrt{2\pi}} [-x \exp(-x^2/2)]_{-\infty}^{+\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp(-x^2/2) dx = 0 + \int_{-\infty}^{+\infty} f(x) dx = 1,$$

puisque f est une densité de probabilité.

$$\begin{aligned} 3. P(X \leq -x) &= \int_{-\infty}^{-x} f(t) dt \\ &= \int_x^{+\infty} f(-s) ds \text{ par le changement de variable } s = -t, \\ &= \int_x^{+\infty} f(s) ds \text{ par parité de } f, \\ &= P(X \geq x). \end{aligned}$$

4.

$$P(-x \leq X \leq x) = P(X \leq x) - P(X \leq -x) = P(X \leq x) - P(X \geq x) = F(x) - (1 - F(x)).$$

■

Définition 5.18 La loi normale de paramètres $\mu \in \mathbb{R}$ et σ^2 (avec $\sigma > 0$) est la loi de densité

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

Si X suit cette loi, on note $X \sim \mathcal{N}(\mu, \sigma^2)$.

Proposition 5.19 Soit $Y \sim \mathcal{N}(\mu; \sigma^2)$. Alors

1. $E[Y] = \mu$;
2. $\text{Var}[Y] = \sigma^2$;
3. $\forall a \in \mathbb{R}^*, \forall b \in \mathbb{R}, aY + b \sim \mathcal{N}(a\mu + b; a^2\sigma^2)$. En particulier

$$\frac{Y - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

Remarque La dernière propriété entraîne aussi que pour tout $\mu \in \mathbb{R}$ et $\sigma > 0$,

$$\text{si } X \sim \mathcal{N}(0, 1), \text{ alors } \sigma X + \mu \sim \mathcal{N}(\mu, \sigma^2).$$

Preuve. On commence par montrer comme dans l'exemple 5.11, que si $X \sim \mathcal{N}(0, 1)$, alors $Y = \sigma X + \mu \sim \mathcal{N}(\mu, \sigma^2)$. On montre en fait ainsi le point 3. de la proposition. Puis les points 1. et 2. en découlent : on a $E(Y) = E(\sigma X + \mu) = \sigma E(X) + \mu = \mu$, puisque $E(X) = 0$. De même $\text{Var}(Y) = \text{Var}(\sigma X + \mu) = \text{Var}(\sigma X) = \sigma^2 \text{Var}(X) = \sigma^2$, puisque $\text{Var}(X) = 1$. ■

Exercice. Soit $X \sim \mathcal{N}(\mu, \sigma^2)$, et soient $a < b$ donnés. Exprimer $P(a < X < b)$ en fonction de la fonction de répartition de la loi normale centrée réduite.

Nous aurons besoin d'un résultat que nous énonçons sans démonstration :

Théorème 5.20 *Si X et Y sont deux variables normales indépendantes, avec $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, alors $X + Y$ est une variable normale $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

Remarque Le seul résultat que nous ne sommes pas capables de démontrer dans le théorème 5.20, est que la somme de variables normales indépendantes est encore une variable normale. Si on admet ce résultat, alors on est capable de trouver les valeurs de μ et σ telles que $X + Y \sim \mathcal{N}(\mu, \sigma^2)$. En effet, on a

$$\mu = E[X + Y] = E[X] + E[Y] = \mu_1 + \mu_2,$$

et

$$\sigma^2 = \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) = \sigma_1^2 + \sigma_2^2,$$

puisque X et Y sont indépendantes.

Chapitre 6

Théorèmes limites.

6.1 Loi des grands nombres.

6.1.1 Motivation.

Supposons que l'on veuille connaître la proportion p de fumeurs dans une population. Une façon naturelle d'avoir une idée de p , est d'interroger un (grand) nombre n de gens dans la population considérée, et de leur demander s'ils sont fumeurs ou pas. Si on note \hat{p}_n la proportion de fumeurs dans l'échantillon de personnes interrogées, il est naturel de penser que $\hat{p}_n \approx p$ quand n est grand. Le théorème qui justifie cette intuition est la célèbre loi des grands nombres. Notons que p est un nombre, alors que \hat{p}_n est une variable aléatoire : si on change l'échantillon de personnes interrogées, la valeur de \hat{p}_n change. Ainsi, on peut se demander quel sens on doit donner à l'affirmation $\hat{p}_n \approx p$? Doit-on comprendre que

1. $\hat{p}_n \approx p$ pour n'importe quel grand échantillon choisi ?
2. si on refaisait l'expérience sur un grand nombre d'échantillons, il y a beaucoup d'échantillons pour lesquels $\hat{p}_n \approx p$?
3. si on refaisait l'expérience sur un grand nombre d'échantillons, la moyenne des proportions \hat{p}_n obtenues sur chaque échantillon, est proche de p ?

Chacune de ces questions correspond à **une** notion de convergence de variables aléatoires, et à **une** version de la loi des grands nombres. Pour être un peu plus précis, modélisons l'expérience précédente. Comme on veut prendre n grand, on considère en théorie qu'on peut interroger un nombre infini de personnes. Ainsi, l'espace des issues est

$$\Omega = \{\omega = (\omega_i, i \in \mathbb{N}^*) : \omega_i \in \{1, 0\}\},$$

où $\omega_i = 1$ signifie que la $i^{\text{ème}}$ personne interrogée est un fumeur.

La probabilité P sur Ω est telle que pour tout n , et tout $(\epsilon_1, \dots, \epsilon_n) \in \{0, 1\}^n$,

$$P(\{\omega : \omega_1 = \epsilon_1, \dots, \omega_n = \epsilon_n\}) = P_1(\epsilon_1) \cdots P_1(\epsilon_n),$$

où $P_1(1) = p$ et $P_1(0) = 1 - p$.

Sur Ω , on définit les variables $X_i : \Omega \rightarrow \{0, 1\}$. Les variables X_i sont indépendantes

$$\omega \mapsto \omega_i$$

et de loi de Bernoulli de paramètre p . On a de plus,

$$\hat{p}_n = \frac{X_1 + \cdots + X_n}{n} = \frac{S_n}{n}.$$

On peut alors reformuler les questions précédentes par :

1. est-ce-que $\forall \omega, \lim_{n \rightarrow \infty} \hat{p}_n(\omega) = p$? (loi forte des grands nombres).
2. si $\epsilon > 0$, est-ce-que $\lim_{n \rightarrow \infty} P[|\hat{p}_n - p| \leq \epsilon] = 1$? (loi faible des grands nombres)
3. est-ce-que $\lim_{n \rightarrow \infty} E[\hat{p}_n] = p$?

On pourrait aussi se demander si $E[|\hat{p}_n - p|]$ tend vers 0 quand $n \rightarrow \infty$, si $E[|\hat{p}_n - p|^2]$ tend vers 0 quand $n \rightarrow \infty$, ...

La réponse à toutes ces questions est affirmative. Nous allons le montrer dans ce qui suit, sauf pour la question 1 (loi forte des grands nombres).

6.1.2 Espérance et variance d'une moyenne empirique.

Soit (Ω, \mathcal{A}, P) un espace de probabilité, et soient $(X_i)_{i \geq 1}$ des v.a. **indépendantes et de même loi**, telles que $E[(X_1)^2]$ ($= E[(X_i)^2]$ pour tout i) est bien définie. On pose

$$S_n \triangleq X_1 + \dots + X_n, \quad \text{et} \quad \bar{X}_n \triangleq \frac{S_n}{n}.$$

\bar{X}_n s'appelle la **moyenne empirique** des X_i .

Proposition 6.1 *On a*

1. $E[\bar{X}_n] = E(X_1)$.
2. $\text{Var}(\bar{X}_n) = \frac{\text{Var}(X_1)}{n}$.

Si on applique cette proposition à l'exemple du sondage fumeur/non fumeur, les X_i sont des variables $\mathcal{B}(p)$, et on a donc $E[X_1] = p$, $\text{Var}(X_1) = p(1-p)$. Par ailleurs $\bar{X}_n = \hat{p}_n$ et la proposition dit que

$$E[\hat{p}_n] = p,$$

$$E[(\hat{p}_n - p)^2] = \text{Var}(\bar{X}_n) = \frac{p(1-p)}{n} \xrightarrow[n \rightarrow \infty]{} 0.$$

Preuve.

1. Par linéarité de l'espérance, $E[\bar{X}_n] = \frac{E[X_1] + \dots + E[X_n]}{n} = \frac{nE[X_1]}{n} = E[X_1]$.
2. $\text{Var}(\bar{X}_n) = \text{Var}\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{\text{Var}(X_1 + \dots + X_n)}{n^2}$
 $= \frac{\text{Var}(X_1) + \dots + \text{Var}(X_n)}{n^2}$ car les X_i sont indépendantes,
 $= \frac{n\text{Var}(X_1)}{n^2}$ car les X_i sont de même loi,
 $= \frac{\text{Var}(X_1)}{n}$.

Remarque Si les X_i ont pour loi la loi normale $\mathcal{N}(\mu, \sigma^2)$, on peut dire beaucoup mieux que la proposition précédente. En fait les résultats du chapitre précédent entraînent que

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

■

6.1.3 Inégalités de Markov et de Chebyshev.

Proposition 6.2 *Inégalité de Markov.*

Soit X une v.a. **positive** telle que $E[X]$ est définie. Alors, pour tout $c > 0$,

$$P(X \geq c) \leq \frac{E[X]}{c}.$$

Preuve. On pose $A = \{X \geq c\}$ et on note que

$$c \mathbb{1}_A(\omega) \leq X(\omega).$$

On prend l'espérance de chaque terme

$$E[c \mathbb{1}_A] = cP(A) \leq E[X],$$

et le résultat en découle. ■

Corollaire 6.3 *Inégalité de Chebyshev.*

Soit X une v.a. de moyenne μ et de variance σ^2 . Pour tout $\epsilon > 0$,

$$P[|X - \mu| \geq \epsilon] \leq \frac{\sigma^2}{\epsilon^2}.$$

Remarque Si on applique l'inégalité de Chebyshev avec $\epsilon = 2\sigma$, on voit que

$$P[|X - \mu| \geq 2\sigma] \leq \frac{1}{4}.$$

Cette borne n'est pas toujours très bonne, puisque si $X \sim \mathcal{N}(0, 1)$, on peut voir sur les tables de la loi normale que $P(|X| \geq 2) \approx 4,5\%$. Le principal avantage (et le principal défaut) de l'inégalité de Chebyshev est qu'elle ne dépend de la loi de X que par l'espérance et la variance de X .

Preuve.

$$\begin{aligned} P[|X - \mu| \geq \epsilon] &= P[(X - \mu)^2 \geq \epsilon^2] \\ &\leq \frac{E[(X - \mu)^2]}{\epsilon^2} \text{ par l'inégalité de Markov appliquée à } Y = (X - \mu)^2, \\ &= \frac{\text{Var}(X)}{\epsilon^2} \text{ car } E[X] = \mu. \end{aligned}$$
■

6.1.4 Loi faible des grands nombres.

Théorème 6.4 Soit $(X_i)_{i \geq 1}$ des v.a. indépendantes et de même loi, de moyenne μ et de variance σ^2 . Pour tout $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0.$$

On dit que \bar{X}_n converge en probabilité vers μ .

Preuve. On applique l'inégalité de Chebyshev à la variable $Y = \bar{X}_n$. On a déjà vu dans la proposition 6.1 que $E(\bar{X}_n) = \mu$, et que $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$. Ainsi,

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\text{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \xrightarrow{n \rightarrow \infty} 0.$$

■

Si on applique ce théorème à l'exemple du sondage fumeur/non fumeur, on obtient que $\lim_{n \rightarrow \infty} P[|\hat{p}_n - p| \geq \epsilon] = 0$.

6.2 Convergence en loi

Définition 6.5 Soient $(X_n)_n$ une suite de v.a. et X une autre v.a. On dira que (X_n) converge en loi vers X si

1. pour tout $x \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} P(X_n \leq x) = P(X \leq x),$$

si X est à densité;

2. pour tout $k \in \mathbb{N}$,

$$\lim_{n \rightarrow \infty} P(X_n = k) = P(X = k),$$

si (X_n) et X sont à valeurs dans \mathbb{N} .

Remarque Cet énoncé porte uniquement sur les lois de X_n et X . En particulier la convergence en loi ne donne aucune information sur la convergence de $X_n(\omega)$ vers $X(\omega)$, à ω fixé. Ainsi on pourra dire que la suite (X_n) converge en loi vers une certaine loi \mathcal{L} , pour dire que si X est une v.a. de loi \mathcal{L} , alors (X_n) converge en loi vers X .

Voyons maintenant quelques exemples classiques.

Proposition 6.6 Si X_n suit la loi uniforme sur $\{0, \dots, n\}$, alors la loi de (X_n/n) tend vers la loi uniforme sur $[0, 1]$.

Preuve. On note $[x]$ la partie entière du réel x , c'est-à-dire l'entier relatif n tel que $n \leq x < n + 1$. Prenons $0 \leq x \leq 1$. On a

$$\{X_n \leq [nx]\} \subset \{X_n/n \leq x\} \subset \{X_n \leq [nx] + 1\}.$$

Comme pour tout k , $P(X_n \leq k) = (k + 1)/(n + 1)$, on obtient :

$$\frac{[nx] + 1}{n + 1} \leq P\left(\frac{X_n}{n} \leq x\right) \leq \frac{[nx] + 2}{n + 1}.$$

Le résultat découle alors du fait que pour tout $x \geq 0$, $[nx]/(n + 1)$ tend vers x , lorsque n tend vers l'infini. ■

Proposition 6.7 Si X_n suit la loi binomiale $\mathcal{B}(n, p_n)$ et si (np_n) tend vers $\lambda > 0$, alors la loi de (X_n) converge vers la loi de Poisson $\mathcal{P}(\lambda)$.

Preuve. On remarque d'abord que puisque np_n tend vers une valeur finie, p_n tend nécessairement vers zéro. Prenons alors $k \in \mathbb{N}$ et $n \geq k$. On a

$$\begin{aligned} P(X_n = k) &= \frac{n!}{k!(n-k)!} p_n^k (1-p_n)^{n-k} \\ &= \frac{\lambda^k}{k!} \left(\frac{np_n}{\lambda}\right)^k \frac{n(n-1)\dots(n-k+1)}{n^k} e^{(n-k)\ln(1-p_n)} \\ &\sim \frac{\lambda^k}{k!} e^{-\lambda}, \end{aligned}$$

car

$$(n-k)\ln(1-p_n) \sim -np_n + kp_n \sim -\lambda.$$

■

Proposition 6.8 Soit $(p_n)_{n \geq 1}$ une suite dans $[0, 1]$, telle que np_n converge vers un certain réel $\lambda > 0$. Pour tout $n \geq 1$, soit T_n une v.a. de loi $\mathcal{G}(p_n)$. Alors (T_n/n) converge en loi vers la loi exponentielle de paramètre λ .

Preuve. La loi de T_n est la loi géométrique de paramètre p_n . Donc

$$\begin{aligned} P(T_n \geq k) &= \sum_{j \geq k} p_n (1-p_n)^{j-1} \\ &= p_n \frac{(1-p_n)^{k-1}}{1-(1-p_n)} = (1-p_n)^{k-1}. \end{aligned}$$

Du coup, pour tout $t > 0$,

$$\begin{aligned} P\left(\frac{T_n}{n} > t\right) &= P(T_n > nt) \\ &\sim (1-p_n)^{nt} = e^{nt \ln(1-p_n)} \\ &\sim e^{-np_n t} \\ &\sim e^{-\lambda t} = \int_t^\infty \lambda e^{-\lambda s} ds. \end{aligned}$$

■

6.3 Théorème central limite.

6.3.1 Énoncé général

La loi des grands nombres nous dit que dans l'exemple du sondage fumeur/non fumeur, $\hat{p}_n \xrightarrow[n \rightarrow \infty]{} p$. Mais à quoi ressemble la déviation $\hat{p}_n - p$ et quelle est sa taille typique ? La réponse est donnée par le théorème central limite.

Théorème 6.9 (Théorème central limite) (*admis*)

Soit $(X_i)_{i \geq 1}$ une suite de v.a. indépendantes de même loi, de moyenne μ et de variance σ^2 . Soit $\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$, et

$$Z_n = \sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right).$$

Alors pour tout $a, b \in \mathbb{R}$, $a < b$,

$$P[a < Z_n < b] \xrightarrow[n \rightarrow \infty]{} P(a < Z < b), \quad \text{où } Z \sim \mathcal{N}(0, 1).$$

Autrement dit Z_n **converge en loi** vers la loi normale centrée réduite.

Remarque On peut réécrire ce théorème en fonction de $S_n = X_1 + \dots + X_n$:

$$P\left[a < \frac{S_n - n\mu}{\sqrt{n}\sigma} < b\right] \xrightarrow[n \rightarrow \infty]{} P(a < Z < B), \quad \text{où } Z \sim \mathcal{N}(0, 1).$$

Intuitivement, on doit comprendre que lorsque n est grand, la loi de S_n “ressemble” à la loi $\mathcal{N}(n\mu, n\sigma^2)$. On a déjà vu ce résultat lorsque les X_i sont des variables gaussiennes (cf théorème 5.20), auquel cas on n’a même pas besoin de supposer n grand. Ce qui est absolument remarquable, c’est que ce résultat reste vrai **quelle que soit la loi des X_i !** C’est ce qui explique l’importance de la loi normale. **La loi normale est la loi limite de la somme d’un grand nombre de phénomènes indépendants.**

6.3.2 Approximations de la loi binomiale

Revenons sur le cas particulier où les X_i sont des variables de Bernoulli de paramètre $p \in]0, 1[$. Dans ce cas le théorème central limite nous dit que lorsque n est suffisamment grand $S_n = X_1 + \dots + X_n$ peut être approchée par la variable gaussienne $\mathcal{N}(np, np(1-p))$. Mais il ne nous dit pas ce qu’il faut entendre par n grand. En fait, la proposition 6.7 nous apprend d’un autre côté que lorsque p devient de plus en plus petit, cette approximation n’est justifiée que pour n de plus en plus grand. Dit autrement, même si n est très grand (en pratique penser à $n \geq 100$), si np reste petit (en pratique penser à np compris entre 1 et 10), alors l’approximation précédente n’est plus justifiée, et S_n ressemble alors plutôt à une loi de Poisson de paramètre np .

En pratique pour les calculs (voir la partie statistique du cours), si $n \geq 50$, et $n \min(p, 1-p) \geq 15$, on utilisera l’approximation par une loi normale, alors que si $n \geq 50$, et $0, 5 \leq np \leq 15$, on utilisera plutôt l’approximation par une loi de Poisson.

Chapitre 7

Concepts de base de la statistique.

7.1 Modèles Statistiques

Le but de la statistique est de traiter des données (en général en grand nombre) en vue d'en tirer une information utile. Ces données proviennent d'un sondage, de mesures, etc..., et sont donc des réalisations d'un phénomène aléatoire.

Exemple 7.1 L'exemple le plus répandu et le plus simple est celui des sondages. Avant une élection où deux candidats A et B se présentent, on cherche à estimer la proportion de gens qui vont voter pour le candidat A ou B . Pour cela on interroge n individus dans la population. On note $x_i = 1$ si la réponse du i -ème individu est A et $x_i = 0$ sinon. On suppose que les individus sont choisis au hasard, si bien que l'on peut considérer que les x_i sont les réalisations de variables aléatoires indépendantes X_i , de loi $\mathcal{B}(p)$, où p est la proportion de gens dans la population qui voteraient pour A si les élections avaient lieu le jour où est effectué le sondage. Au vu des réalisations des variables X_i on cherche à estimer la valeur du paramètre p , et à tester si elle est supérieure à $1/2$ ou non (c'est-à-dire si le candidat A va gagner ou non l'élection).

Exemple 7.2 Un chimiste veut connaître la concentration μ en acétone d'un certain produit. Pour cela, il prélève dix échantillons de ce produit, dont il détermine la concentration. Il obtient ainsi dix mesures (x_1, \dots, x_{10}) probablement différentes de μ , chaque dosage étant entaché d'erreurs. Il a donc observé le vecteur aléatoire (X_1, \dots, X_{10}) avec $X_i = \mu + \epsilon_i$, où ϵ_i représente les erreurs de dosage. Il dispose des informations a priori suivantes :

- la loi de chaque ϵ_i est indépendante de μ (les erreurs de dosages ne dépendent pas de la concentration du produit) ;
- les variables aléatoires ϵ_i sont indépendantes (les erreurs de mesure faites lors d'un dosage ne se répercutent pas sur les dosages suivants) ;
- les ϵ_i ont même loi (la méthode de dosage utilisée est toujours la même) ;
- la moyenne des ϵ_i est nulle (la méthode de dosage n'introduit pas une erreur systématique) ;
- la variance σ^2 des ϵ_i est la précision de la méthode de dosage utilisée. Elle peut être connue ou non, selon les cas.

Il fait en outre l'hypothèse que la loi commune des ϵ_i est la loi normale (de moyenne 0, et de variance σ^2). Cette hypothèse peut se justifier par le TCL (théorème central limite), si l'on considère que les erreurs de dosage proviennent de l'accumulation d'un grand nombre

de petites erreurs indépendantes. La loi des X_i est donc la loi normale de moyenne μ et variance σ^2 .

La démarche de la statistique est donc en quelque sorte inverse de celle des probabilités. En probabilité, on se donne une variable X de loi connue P , et on tire de ce modèle des informations sur les réalisations de cette variable (observation la plus probable, probabilité pour qu’une observation soit dans un certain ensemble, etc...)

En statistique, la donnée de base est la **donnée d’une réalisation** $x = X(\omega)$ d’une variable aléatoire X de **loi inconnue** P . On fera par ailleurs souvent l’hypothèse a priori que la loi P appartient à un certain ensemble de probabilités $(P_\theta, \theta \in \Theta)$. Dans l’exemple 7.1 la loi inconnue appartient à la famille $(\mathcal{B}(p), p \in [0, 1])$. Dans l’exemple 7.2 la loi inconnue appartient à la famille $(\mathcal{N}(\mu, \sigma^2), \mu \in [0, 1], \sigma^2 > 0)$, etc....

La famille des lois $\{P_\theta; \theta \in \Theta\}$ est ce que l’on appelle le **modèle statistique**. Celui-ci décrit l’information dont on dispose a priori sur le phénomène aléatoire considéré. Θ est l’ensemble des **paramètres inconnus**. Il décrit tout ce qu’on ignore sur le phénomène aléatoire. On utilisera par ailleurs la définition suivante :

Définition 7.3 *On dit que (X_1, \dots, X_n) est un **n-échantillon** de loi \mathcal{L} si les variables X_i sont indépendantes et de même loi \mathcal{L} .*

7.2 Quel genre d’informations tirer d’un modèle statistique ?

Une fois le modèle choisi, on peut se poser différentes questions quant à la vraie probabilité. Par exemple dans l’exemple 7.2, on peut vouloir estimer le paramètre μ , ou au moins donner un intervalle dans lequel il a toute chance de se situer. On peut aussi simplement se demander s’il est supérieur ou non à une certaine valeur μ_0 . Ainsi dans l’exemple 7.1, on se demandera en général simplement si p est supérieur ou non à $1/2$.

Le premier type de questions relève de la **théorie de l’estimation**. Il s’agit d’estimer des caractéristiques numériques du modèle. La réponse peut se faire soit sous la forme d’une valeur numérique (“la concentration est proche de 10%”) et on parle alors d’**estimation ponctuelle**; soit sous la forme d’un intervalle (“la concentration se situe entre 1% et 20%”), et on parle alors d’**estimation par intervalle**.

Le deuxième type de questions requiert la réponse ”oui” ou ”non”. Il s’agit de décider si le paramètre qui nous intéresse, appartient à un sous-ensemble donné Θ_1 de Θ , ou pas. On parle alors de **tests d’hypothèses**, ou encore de **théorie de la décision**.

7.3 Estimation

7.3.1 Estimateur.

Reprenons l’exemple 7.1, et notons

$$X_i = \begin{cases} 1 & \text{si la } i\text{-ème personne interrogée déclare vouloir voter pour le candidat } A, \\ 0 & \text{sinon;} \end{cases}$$

$$S_n = \sum_{i=1}^n X_i = \text{nombre de personnes interrogées déclarant vouloir voter pour } A.$$

Les variables X_i sont indépendantes de loi $\mathcal{B}(p)$.

Une estimation naturelle de la vraie proportion p de personnes voulant voter A est donnée par la proportion empirique $\bar{X}_n = \frac{1}{n}S_n =$ proportion de personnes déclarant vouloir voter pour le candidat A . Cette démarche est justifiée par la loi des grands nombres qui assure que dans le modèle choisi, $\lim_{n \rightarrow \infty} \bar{X}_n = E_p(X_1) = p$.

Définition 7.4 Une **observation** X est une v.a. (ou un échantillon de v.a.) de loi appartenant à la famille $(P_\theta; \theta \in \Theta)$. On appelle **estimateur** du paramètre θ toute fonction $\hat{\theta} = h(X)$ de l'observation X .

Remarque Il faut ici comprendre qu'un estimateur de θ doit pouvoir être calculé dès qu'on dispose de l'échantillon de données. En aucun cas, un estimateur ne peut dépendre de paramètres inconnus.

Définition 7.5 Un estimateur $\hat{\theta}_n$ (où n est la taille de l'échantillon de données) est dit **consistant** si $\hat{\theta}_n \xrightarrow[n \rightarrow \infty]{} \theta$.

7.3.2 Construction d'estimateurs

Méthode des moments

Dans l'exemple 7.1, nous avons choisi pour estimateur de $p = E(X_1)$, la moyenne empirique $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$, qui est un estimateur consistant, d'après la loi des grands nombres. Cette façon de procéder peut se généraliser dans un cadre très général.

Supposons qu'on ait observé un n -échantillon (X_1, \dots, X_n) de loi P_θ et que le paramètre θ à estimer, soit tel que pour une certaine fonction f , on ait

$$\theta = E[f(X_1)].$$

La loi des grands nombres nous dit que $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n f(X_i)$ est un estimateur consistant de θ . On dit qu'on a estimé θ par la **méthode des moments**. Cette méthode consiste à exprimer le paramètre θ comme l'espérance d'une fonction de X_1 ("un moment"), et à remplacer l'espérance par une moyenne empirique.

Ainsi dans l'exemple 7.2, le chimiste observe $(X_1(\omega), \dots, X_n(\omega))$, une réalisation d'un n -échantillon de la loi $\mathcal{N}(\mu, \sigma^2)$, où μ et σ^2 sont les paramètres à estimer. Or $\mu = E(X_1)$, et un estimateur de μ par la méthode des moments est donc

$$\hat{\mu}_n = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

De même pour estimer σ^2 , on estime $E(X_1^2)$ par la méthode des moments, et on trouve donc l'estimateur

$$\begin{aligned} \hat{\sigma}_n^2 &= \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2. \end{aligned}$$

Méthode du maximum de vraisemblance

Une autre méthode couramment employée est l'**estimation par maximum de vraisemblance**. Elle consiste à estimer le paramètre, par la valeur qui rend l'observation faite la plus probable.

Lois discrètes.

Exemple 7.6 Reprenons l'exemple 7.1, et supposons que sur 1000 personnes interrogées, 450 déclarent vouloir voter A . Soit S_n la v.a. représentant le nombre de personnes déclarant voter A , sur $n = 1000$ interrogées. Sur notre échantillon, on a observé $S_n(\omega) = 450$. Cherchons la valeur de p qui rend cette observation la plus probable. Comme $S_n \sim \mathcal{B}(1000, p)$,

$$P[S_n = 450] = C_{1000}^{450} p^{450} (1-p)^{1000-450}.$$

La valeur de p cherchée maximise donc la fonction $h : p \mapsto p^{450}(1-p)^{550}$. Pour trouver cette valeur, on regarde en quel point la dérivée de la fonction h s'annule, et on trouve le point $\hat{p}_n(\omega) = \frac{450}{1000} = \frac{S_n(\omega)}{n}$. L'estimateur du maximum de vraisemblance est alors $\hat{p}_n = S_n/1000$. Il coïncide donc ici avec l'estimateur obtenu par la méthode des moments.

Exemple 7.7 Soit N le nombre de voitures qui prennent l'autoroute Aix-Marseille entre 10h et 11h. On a observé $N(\omega) = 90$. On suppose que N suit une loi de Poisson de paramètre λ . Quel est l'estimateur du maximum de vraisemblance de λ ?

On a $P(N = 90) = \exp(-\lambda) \frac{\lambda^{90}}{90!}$. Il s'agit ici de trouver λ qui maximise la fonction $h : \lambda \mapsto \exp(-\lambda) \lambda^{90}$. Pour trouver cette valeur, on regarde en quel point la dérivée de h s'annule, et on trouve le point $\hat{\lambda}(\omega) = 90 = N(\omega)$. L'estimateur du maximum de vraisemblance est alors $\hat{\lambda} = N$. Noter que $E_\lambda(N) = \lambda$, et donc à nouveau, cet estimateur coïncide avec un estimateur par la méthode des moments.

Lois à densité.

Dans le cas où les observations faites ($X_1(\omega) = x_1, \dots, X_n(\omega) = x_n$) sont une réalisation d'un n -échantillon d'une variable à densité, quelles que soient les valeurs x_1, \dots, x_n observées, on a toujours $P_\theta[X_1 = x_1, \dots, X_n = x_n] = 0$. Il faut donc modifier légèrement notre façon de mesurer la "probabilité" d'une observation. Ainsi, si $f_\theta(x)$ désigne la densité de la loi de X_1 , on a

$$P_\theta[X_1 \in [x_1, x_1 + dx_1], \dots, X_n \in [x_n, x_n + dx_n]] \approx f_\theta(x_1) \cdots f_\theta(x_n) dx_1 \cdots dx_n,$$

Cette fois, l'estimateur du maximum de vraisemblance est la valeur du paramètre qui maximise $\theta \mapsto f_\theta(x_1) \cdots f_\theta(x_n)$, où (x_1, \dots, x_n) sont les valeurs observées.

Définition 7.8 Si l'observation est constituée d'un n -échantillon X_1, \dots, X_n de variables de densité f_θ , on appelle **vraisemblance** du modèle, la variable aléatoire

$$V_\theta(X_1, \dots, X_n) = f_\theta(X_1) \cdots f_\theta(X_n).$$

On appelle **estimateur du maximum de vraisemblance** toute valeur du paramètre θ qui maximise la fonction $\theta \mapsto V_\theta(X_1, \dots, X_n)$.

Exemple 7.9 Dans l'exemple 7.2, le chimiste observe $(X_1(\omega), \dots, X_n(\omega))$, une réalisation d'un n -échantillon de la loi $\mathcal{N}(\mu, \sigma^2)$, où la concentration μ en acétone et la précision σ^2 de la méthode de dosage sont les paramètres à estimer. Dans ce modèle, la vraisemblance est

$$V_{\mu, \sigma^2}(X_1, \dots, X_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi\sigma^2}^n} \exp\left(-\frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2}\right).$$

Pour trouver les estimateurs du maximum de vraisemblance de μ et σ^2 , il faut maximiser la fonction $V_{\mu, \sigma^2}(X_1, \dots, X_n)$ en μ et σ^2 . La maximisation en μ revient à minimiser en μ la fonction $\mu \mapsto \sum_{i=1}^n (X_i - \mu)^2$. En notant $\hat{\mu}_n$ l'estimateur du maximum de vraisemblance pour μ , l'annulation de la dérivée conduit à l'équation :

$$2 \sum_{i=1}^n (X_i - \hat{\mu}_n) = 0 \iff \hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n,$$

qui correspond bien à un minimum. Si $\hat{\sigma}_n^2$ est l'estimateur du maximum de vraisemblance pour σ^2 , il doit donc maximiser la fonction $\sigma^2 \mapsto \frac{1}{\sqrt{2\pi\sigma^2}^n} \exp\left(-\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{2\sigma^2}\right)$. L'annulation de la dérivée conduit à l'équation :

$$-\frac{n}{2\hat{\sigma}_n^2} + \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{2\hat{\sigma}_n^4} = 0 \iff \hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Les estimateurs du maximum de vraisemblance pour la moyenne et la variance dans les échantillons gaussiens redonnent donc les estimateurs empiriques (qui sont ceux obtenus par la méthode des moments).

Exemple 7.10 Supposons que (X_1, \dots, X_n) soit un n -échantillon de loi $\mathcal{E}(1/\lambda)$, où $\lambda > 0$ est un paramètre inconnu. Dans ce cas pour trouver l'estimateur du maximum de vraisemblance, il faut maximiser la fonction $h : \lambda \mapsto \frac{1}{\lambda^n} \exp(-(X_1 + \dots + X_n)/\lambda)$. On obtient l'estimateur $\hat{\lambda}_n = \bar{X}_n$, qui est là encore le même que celui obtenu par la méthode des moments. On observera au passage que si la loi commune des X_i est plutôt $\mathcal{E}(\lambda)$, alors l'estimateur du maximum de vraisemblance est $\hat{\lambda}_n = 1/\bar{X}_n$.

Exemple 7.11 Un exemple classique où les deux méthodes précédentes ne donnent pas le même estimateur, est celui où (X_1, \dots, X_n) est un n -échantillon de loi uniforme $\mathcal{U}([0, \theta])$, avec $\theta > 0$ inconnu. En effet dans ce cas l'estimateur donné par la méthode du maximum de vraisemblance est (exercice !)

$$\hat{\theta}_n = \max(X_1, \dots, X_n),$$

alors que celui obtenu par la méthode des moments est $2\bar{X}_n$.

7.3.3 Biais et risque d'un estimateur

Définition 7.12 Le **biais** d'un estimateur $\hat{\theta}$ est la fonction $b : \theta \mapsto E(\hat{\theta}) - \theta$.

Un estimateur $\hat{\theta}$ est dit **sans biais**, si pour tout $\theta \in \Theta$, $E(\hat{\theta}) = \theta$.

Ainsi dans les exemples 7.1 et 7.2, l'estimateur \bar{X}_n est sans biais. Par contre l'estimateur de la variance $\hat{\sigma}_n^2$ construit dans l'exemple 7.9 n'est pas sans biais, puisque $E(\hat{\sigma}_n^2) = (n-1)\sigma^2/n$. En revanche l'estimateur suivant est sans biais :

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

Définition 7.13 On appelle **risque** (quadratique) d'un estimateur $\hat{\theta}$ la fonction qui à tout $\theta \in \Theta$ associe la quantité :

$$R(\theta) = E[(\hat{\theta} - \theta)^2].$$

On a alors la **décomposition biais-variance** :

$$R(\theta) = b(\theta)^2 + \text{Var}(\hat{\theta}),$$

où $b(\theta) = E(\hat{\theta}) - \theta$ est le biais, et $\text{Var}(\hat{\theta}) = E(\hat{\theta}^2) - E(\hat{\theta})^2$ est la variance de $\hat{\theta}$. Donc pour un estimateur sans biais, le risque et la variance coïncident. Par exemple si (X_1, \dots, X_n) est un n -échantillon de loi inconnue et $\theta = E(X_1)$, alors on a vu que \bar{X}_n est un estimateur consistant de θ , et sans biais. Son risque est donc égal à sa variance, soit $\text{Var}(X_1)/n$.

Un principe important est le suivant :

Entre deux estimateurs d'un même paramètre, le meilleur est celui qui minimise le risque.

Attention toutefois, **les meilleurs estimateurs ne sont pas toujours les estimateurs sans biais**. Ainsi dans l'exemple 7.11, on peut montrer que $\hat{\theta}_n$ a un risque qui est de l'ordre de $1/n^2$. En revanche, on a vu que la variance de \bar{X}_n vaut σ^2/n . Donc lorsque n est grand, $\hat{\theta}_n$ est un meilleur estimateur, même s'il est biaisé. De même si (X_1, \dots, X_n) est un n -échantillon de loi $\mathcal{N}(\mu, \sigma^2)$, l'estimateur $\hat{\Sigma}_n^2$ a un plus grand risque que $\hat{\sigma}_n^2$ (voir plus haut).

Chapitre 8

Intervalles de confiance

8.1 Echantillon de Bernoulli

On veut se faire une idée de la proportion p d'ampoules défectueuses d'un fabricant. On prélève pour cela un échantillon de 1000 ampoules au hasard. On note X_1, \dots, X_{1000} les v.a. représentant les résultats obtenus pour les 1000 ampoules, et comme dans l'exemple 7.1, on fait l'hypothèse que (X_1, \dots, X_{1000}) est un 1000-échantillon de la loi $\mathcal{B}(p)$. Dans ce cas, on a vu que \bar{X}_n est un estimateur consistant de p . Si sur 1000 ampoules testées, 90 sont défectueuses, on a donc de bonnes raisons de penser que $p \approx 9\%$. Mais quelle est la précision de cette estimation ? Une réponse est donnée par les intervalles de confiance, les fameuses "fourchettes" des sondages électoraux. Il s'agit ici de dire que le paramètre p est dans un intervalle $[\bar{X}_n - t, \bar{X}_n + t]$ avec grande probabilité, 95% par exemple. On cherche donc t tel que

$$\begin{aligned} P_p [p \in [\bar{X}_n - t, \bar{X}_n + t]] = 95\% &\Leftrightarrow P_p [|\bar{X}_n - p| \leq t] = 95\% \\ &\Leftrightarrow P_p \left[\frac{|\bar{X}_n - p|}{\sqrt{p(1-p)/n}} \leq \frac{t}{\sqrt{p(1-p)/n}} \right] = 95\%. \end{aligned}$$

Si n est suffisamment grand, le TLC nous dit que $\frac{|\bar{X}_n - p|}{\sqrt{p(1-p)/n}}$ est approximativement de loi $\mathcal{N}(0, 1)$. Soit alors $Z \sim \mathcal{N}(0, 1)$, et soit z l'unique valeur telle que

$$P[|Z| \leq z] = 95\%.$$

Les tables de la loi normale donnent $z \approx 1.96$. Ainsi, on a

$$\frac{t}{\sqrt{p(1-p)/n}} \approx 1.96.$$

Cette équation ne nous permet pas de déterminer t puisqu'elle contient deux inconnues, t et p . Mais si n est grand, p peut être approché par \bar{X}_n . On arrive alors à

$$t \approx 1.96 \sqrt{\bar{X}_n(1 - \bar{X}_n)/n}.$$

On peut donc écrire

$$P_p \left(p \in \left[\bar{X}_n - 1,96 \frac{\sqrt{\bar{X}_n(1 - \bar{X}_n)}}{\sqrt{n}}, \bar{X}_n + 1,96 \frac{\sqrt{\bar{X}_n(1 - \bar{X}_n)}}{\sqrt{n}} \right] \right) \approx 95\%.$$

On dit que l'intervalle

$$I = \left[\bar{X}_n - 1,96 \frac{\sqrt{\bar{X}_n(1 - \bar{X}_n)}}{\sqrt{n}}, \bar{X}_n + 1,96 \frac{\sqrt{\bar{X}_n(1 - \bar{X}_n)}}{\sqrt{n}} \right]$$

est un intervalle de confiance pour p au niveau de confiance 95%. On notera que les bornes de cet intervalle sont des variables aléatoires. Maintenant au vu des données on observe que

$$t \approx 1.96 \sqrt{0.09(0.91)/1000} = 0.017.$$

On peut donc "parier" au vu des données que $p \in [0.09 - 0.017; 0.09 + 0.017]$. Ce qu'il faut comprendre ici, c'est que si l'on refait la même expérience (compter le nombre d'ampoules défectueuses sur 1000 testées) un grand nombre de fois, 95% des intervalles de confiance obtenus contiennent la vraie valeur de p . Il s'agit donc bien d'un pari que l'on fait sur notre échantillon, mais il est tout à fait possible que nous n'ayons pas de chance, et que notre échantillon de données soit parmi les 5% de "mauvais" échantillons.

8.2 Définitions générales

Définition 8.1 Soit (X_1, \dots, X_n) un n -échantillon de loi P_θ . Un **intervalle de confiance au niveau de sécurité** $1 - \alpha$ (α petit), est un intervalle $[i(X_1, \dots, X_n); I(X_1, \dots, X_n)]$ dont les bornes sont des variables aléatoires qui ne dépendent que de l'observation, et tel que $\forall \theta$,

$$P_\theta [\theta \in [i(X_1, \dots, X_n); I(X_1, \dots, X_n)]] \geq 1 - \alpha.$$

L'intervalle $[i(X_1(\omega), \dots, X_n(\omega)); I(X_1(\omega), \dots, X_n(\omega))]$, associé à une expérience ω est appelé **version observée** de l'intervalle.

En général, la construction d'un intervalle de confiance est basée sur un estimateur $\hat{\theta}_n$ de θ (\bar{X}_n dans l'exemple précédent). On cherche ensuite une fonction $F(\hat{\theta}_n, \theta)$ de l'estimateur $\hat{\theta}_n$ et du paramètre θ , telle que la variable $Z = F(\hat{\theta}_n, \theta)$ **a une loi qui ne dépend plus de θ** . Une telle variable s'appelle un **pivot**. Dans l'exemple, on a pris $Z = \frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}}$ qui est à peu près de loi $\mathcal{N}(0, 1)$ lorsque n est grand. Si $Z = F(\hat{\theta}_n, \theta)$ est un pivot, il est possible de trouver des nombres t_1 et t_2 tels que $P[Z \in [t_1; t_2]] = 95\%$. Il y a même une infinité de solutions pour (t_1, t_2) . Dans l'exemple, on a choisi $t_2 = -t_1$, et $P(|Z| \leq t_1) = 95\%$, avec $Z \sim \mathcal{N}(0, 1)$. On a alors une probabilité de 95% pour que

$$t_1 \leq F(\hat{\theta}_n, \theta) \leq t_2.$$

L'inversion de cette inégalité en θ donne un intervalle de confiance pour θ .

Remarque Dans l'exemple précédent on a utilisé une approximation de la loi du pivot par la loi $\mathcal{N}(0, 1)$. L'intervalle de confiance n'est donc valide que pour de grands échantillons (pour n grand). On dit alors que l'on a un **intervalle de confiance asymptotique**. Une autre manière d'obtenir un intervalle de confiance valable pour tout n , que l'on dit alors **exact**, est d'utiliser l'inégalité de Chebychev. En effet cela permet d'écrire que

$$P(|\bar{X}_n - p| > t) \leq \frac{\text{Var}(\bar{X}_n)}{t^2} = \frac{p(1-p)}{nt^2} \leq \frac{1}{4nt^2},$$

en utilisant que la fonction $p \mapsto p(1-p)$ atteint son maximum en $p = 1/2$. Il suffit ensuite de choisir t tel que $1/4nt^2 = \alpha$, et on obtient alors l'intervalle de confiance

$$J = \left[\bar{X}_n - \frac{1}{2\sqrt{n\alpha}}, \bar{X}_n + \frac{1}{2\sqrt{n\alpha}} \right].$$

Toutefois pour de grandes valeurs de n , on a intérêt à utiliser l'approximation par la loi normale, qui donne des intervalles de confiance de longueur plus petite. En effet la longueur de J est $1/\sqrt{n\alpha}$, alors que celle de I obtenue dans l'exemple précédent est au plus de $1,96/(2\sqrt{n})$ (toujours en utilisant que $\bar{X}_n(1-\bar{X}_n) \leq 1/4$).

8.3 Échantillons gaussiens

On suppose ici que l'observation est constituée d'un n -échantillon (X_1, \dots, X_n) de loi $\mathcal{N}(\mu, \sigma^2)$. Suivant les cas, on supposera σ connu ou inconnu.

8.3.1 Estimateurs de la moyenne et de la variance.

Estimateur de la moyenne.

Posons comme d'habitude

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

\bar{X}_n est la **moyenne empirique** de l'échantillon.

Proposition 8.2 \bar{X}_n est un estimateur consistant de μ . Et on a

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{et} \quad \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Preuve. Par la loi des grands nombres, \bar{X}_n est un estimateur consistant de μ . Par ailleurs, on a déjà vu (cf théorème 5.20) que la somme de variables normales indépendantes est une variable normale. On en déduit que $X_1 + \dots + X_n$ suit la loi normale de moyenne $m = E(X_1 + \dots + X_n) = n\mu$, et de variance $\text{Var}(X_1 + \dots + X_n) = n\sigma^2$, puisque les X_i sont indépendantes. Cela donne le résultat (cf 3. de la proposition 5.19). ■

Ainsi,

si σ est connu, $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ est un pivot pour l'estimation de μ .

Estimateur de la variance.

Définition 8.3 (Loi du chi-deux) Soit (X_1, \dots, X_n) un n -échantillon de loi $\mathcal{N}(0, 1)$. Alors la v.a. $Z = X_1^2 + \dots + X_n^2$ est à densité, et sa loi est appelée loi du chi-deux à n degrés de liberté. On note $Z \sim \chi^2(n)$.

On définit maintenant

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

S_n^2 s'appelle la **variance empirique** de l'échantillon.

Proposition 8.4 *La v.a. S_n^2 est un estimateur consistant de σ^2 , et (admis)*

$$\frac{nS_n^2}{\sigma^2} \sim \chi^2(n-1).$$

Preuve. En développant le carré, il est facile de voir que

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

La loi des grands nombres nous dit que $\frac{1}{n} \sum_{i=1}^n X_i^2$ converge vers $E(X_1^2)$, et que \bar{X}_n converge vers $E(X_1)$. Ainsi S_n^2 converge vers $E(X_1^2) - E(X_1)^2 = \text{Var}(X_1) = \sigma^2$. ■

Ainsi,

la v.a. $\frac{nS_n^2}{\sigma^2}$ est un pivot pour l'estimation de σ^2 .

8.3.2 Intervalle de confiance pour μ avec σ connu.

Il est basé sur l'estimateur \bar{X}_n , et le pivot $Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$. Pour $\alpha > 0$ fixé (petit), soit t tel que $P[|Z| \leq t] = 1 - \alpha$ (t se lit dans les tables de la loi normale). On a donc

$$P\left(\mu \in \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}}t; \bar{X}_n + \frac{\sigma}{\sqrt{n}}t\right]\right) = 1 - \alpha.$$

σ étant connu, les bornes de l'intervalle ne dépendent bien que de l'observation.

8.3.3 Intervalle de confiance pour μ avec σ inconnu.

Définition 8.5 (Loi de Student) *Soit X une v.a. de loi $\mathcal{N}(0, 1)$ et Y une v.a. de loi $\chi^2(n)$ indépendante de X . Alors $Z = \frac{X}{\sqrt{Y}}$ est une v.a. à densité et sa loi est appelée loi de Student à n degrés de liberté. On note $Z \sim \mathcal{S}(n)$.*

L'intervalle de confiance précédemment construit ne convient plus, puisqu'il fait apparaître l'inconnue σ . On utilise alors le théorème suivant :

Théorème 8.6 *(admis) Les v. a. \bar{X}_n et S_n^2 sont indépendantes. On a donc*

$$T_n = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n-1}} \sim \mathcal{S}(n-1).$$

Autrement dit,

si σ est inconnu, T_n est un pivot pour l'estimation de μ .

Ainsi pour α donné (et petit), on cherche t tel que

$$1 - \alpha = P(|T_n| \leq t)$$

à l'aide de la table de Student. On a donc

$$P\left(\mu \in \left[\bar{X}_n - \frac{S_n}{\sqrt{n-1}}t; \bar{X}_n + \frac{S_n}{\sqrt{n-1}}t\right]\right) = 1 - \alpha.$$

Les bornes de l'intervalle ne dépendent bien que de l'observation.

8.3.4 Intervalle de confiance pour σ^2 .

Il est basé sur l'estimateur S_n^2 de σ^2 , et le pivot $nS_n^2/\sigma^2 \sim \chi^2(n-1)$.
On va donc chercher un intervalle $[t_1, t_2]$ tel que pour α petit donné

$$1 - \alpha = P(t_1 \leq \chi^2(n-1) \leq t_2).$$

On peut par exemple faire le choix suivant

$$\alpha/2 = P(\chi^2(n-1) \leq t_1), \quad \text{et} \quad \alpha/2 = P(\chi^2(n-1) \geq t_2).$$

Une lecture des tables de lois du χ^2 donne alors t_1 et t_2 . On a donc

$$P\left(\sigma^2 \in \left[\frac{nS_n^2}{t_2}; \frac{nS_n^2}{t_1}\right]\right) = 1 - \alpha.$$

Les bornes de l'intervalle ne dépendent bien que de l'observation.

Chapitre 9

Premiers tests : échantillons de Bernoulli et échantillons gaussiens

9.1 Échantillons de Bernoulli

Le producteur d'ampoules affirme à ses clients que la proportion p d'ampoules défectueuses est inférieure à 10%. Il veut s'assurer que cela est bien le cas, pour éviter que les lots vendus ne lui soient renvoyés. Avec beaucoup de bon sens, il se dit que si \bar{X}_n est "très grand", p a peu de chances d'être inférieure à 10%. Il va donc décider

- que $p > 10\%$ si $\bar{X}_n > t$;
- que $p \leq 10\%$ si $\bar{X}_n \leq t$;

où t est un nombre à déterminer. **Comment choisir t ?**

Une première attitude peut être de prendre $t = 10\%$.

- Si $\bar{X}_n > 10\%$, je décide que $p > 10\%$.
- Si $\bar{X}_n \leq 10\%$, je décide que $p \leq 10\%$.

Calculons les probabilités de se tromper si on adopte cette règle. Une première façon de se tromper est de décider que $p > 10\%$, alors que ce n'est pas le cas. La probabilité d'erreur associée est au pire des cas,

$$\alpha_1 = \sup_{p \leq 10\%} P_p [\bar{X}_n > 10\%] = \sup_{p \leq 10\%} P_p \left[\frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}} > \frac{10\% - p}{\sqrt{p(1-p)/n}} \right].$$

Si n est grand, on sait par le TLC que $\frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}}$ suit approximativement la loi $\mathcal{N}(0, 1)$, et on a donc

$$\alpha_1 \approx \sup_{p \leq 10\%} P \left[Z > \frac{10\% - p}{\sqrt{p(1-p)/n}} \right], \quad Z \sim \mathcal{N}(0, 1).$$

On peut vérifier que la fonction $p \in [0, 10\%] \mapsto \frac{10\% - p}{\sqrt{p(1-p)/n}}$ est décroissante, et donc que la fonction $p \in [0, 10\%] \mapsto P \left[Z > \frac{10\% - p}{\sqrt{p(1-p)/n}} \right]$ est croissante. Ainsi, le "sup" dans l'expression de α_1 est atteint pour $p = 10\%$. On obtient donc

$$\alpha_1 \approx P[Z > 0] = 50\%.$$

On a donc au pire des cas (qui correspond à $p = 10\%$), 50 % de chances de se tromper en décidant que $p > 10\%$ si on adopte cette stratégie de décision.

L'autre erreur possible est de décider que $p \leq 10\%$, alors que ce n'est pas le cas :

$$\begin{aligned}\alpha_2 &= \sup_{p > 10\%} P_p [\bar{X}_n \leq 10\%] \\ &= \sup_{p > 10\%} P_p \left[\frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}} \leq \frac{10\% - p}{\sqrt{p(1-p)/n}} \right] \\ &\approx \sup_{p > 10\%} P \left[Z \leq \frac{10\% - p}{\sqrt{p(1-p)/n}} \right], \quad Z \sim \mathcal{N}(0, 1).\end{aligned}$$

À nouveau, on peut vérifier que la fonction $p \in]10\%, 1] \mapsto \frac{10\% - p}{\sqrt{p(1-p)/n}}$ est décroissante, et donc que la fonction $p \in]10\%, 1] \mapsto P \left[Z < \frac{10\% - p}{\sqrt{p(1-p)/n}} \right]$ est décroissante. Ainsi, le "sup" dans l'expression de α_2 est obtenu en faisant $p \rightarrow 10\%$. On a donc

$$\alpha_2 \approx P[Z < 0] = 50\%.$$

Dans le pire des cas (i.e. p proche de 10% par valeurs supérieures), on a donc aussi 50 % de chances de se tromper en décidant que $p \leq 10\%$ avec la règle de décision précédente.

Comment améliorer notre règle de décision ? Au lieu de prendre $t = 10\%$, essayons d'ajuster t de façon à réduire les probabilités d'erreur précédentes. Si on reprend les calculs précédents, on a maintenant

$$\begin{aligned}\alpha_1 &= \sup_{p \leq 10\%} P_p [\bar{X}_n > t] \approx \sup_{p \leq 10\%} P \left[Z > \frac{t - p}{\sqrt{p(1-p)/n}} \right], \quad Z \sim \mathcal{N}(0, 1). \\ \alpha_2 &= \sup_{p > 10\%} P_p [\bar{X}_n \leq t] \approx \sup_{p > 10\%} P \left[Z \leq \frac{t - p}{\sqrt{p(1-p)/n}} \right], \quad Z \sim \mathcal{N}(0, 1).\end{aligned}$$

Il est clair sur ces expressions que plus on choisit t grand, plus α_1 devient petit. En revanche, plus t est grand, plus α_2 est grand. **Les deux erreurs ne peuvent pas en général être petites en même temps.**

La méthode utilisée est alors

- de choisir parmi les deux types d'erreurs possibles, l'erreur que l'on veut absolument contrôler (parce que par exemple, elle a des conséquences plus coûteuses que l'autre) ;
- de lui assigner un niveau α petit ;
- de choisir t en conséquence.

Par exemple, si le fabricant d'ampoules préfère se tromper en concluant que $p \geq 10\%$, plutôt qu'en concluant que $p < 10\%$ (auquel cas beaucoup de lots vont lui être renvoyés), il va fixer un niveau α (petit, par exemple 5%) pour la probabilité de conclure que $p < 10\%$, alors que ce n'est pas le cas. Il va donc choisir t tel que

$$\sup_{p \geq 10\%} P_p(\bar{X}_n < t) = \alpha.$$

En pratique, cela revient à privilégier une hypothèse par rapport à l'autre. Le fabricant va tester (H_0) : “ $p \geq 10\%$ ”, contre (H_1) : “ $p < 10\%$ ”.

Par **définition de (H_0)** (et donc de (H_1)), l'erreur qu'il contrôle est

$$\alpha = P_{(H_0)}(\text{Rejeter } (H_0)).$$

α s'appelle le **niveau du test**, ou **l'erreur de première espèce**.

Sa règle de décision va être la suivante :

$$\begin{aligned} \text{Si } \bar{X}_n < t & \text{ Rejet de } (H_0) , \\ \text{Si } \bar{X}_n \geq t & \text{ Acceptation de } (H_0) , \end{aligned}$$

où t est choisi de telle sorte que

$$\alpha = P_{(H_0)}(\text{Rejeter } (H_0)) = \sup_{p \geq 10\%} P_p(\bar{X}_n < t). \quad (9.1)$$

Si on prend $\alpha = 5\%$, le calcul de t (i.e. la résolution de l'équation (9.1)) se fait de la façon suivante :

$$\begin{aligned} 5\% &= \sup_{p \geq 10\%} P_p(\bar{X}_n < t) \\ &= \sup_{p \geq 10\%} P_p \left[\frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}} < \frac{t - p}{\sqrt{p(1-p)/n}} \right] \\ &\approx \sup_{p \geq 10\%} P \left[Z < \frac{t - p}{\sqrt{p(1-p)/n}} \right], \quad Z \sim \mathcal{N}(0, 1), \\ &= P \left[Z < \frac{t - 0.1}{\sqrt{0.1(0.9)/n}} \right] \quad (\text{car le sup est atteint pour } p = 10\%) \end{aligned}$$

La table de la loi normale donne $\frac{t-0.1}{\sqrt{0.1(0.9)/n}} = -1.645$, d'où on déduit la valeur de t

$$t = 0.1 - 1.645 \sqrt{\frac{(0.1)(0.9)}{n}}.$$

Une fois le niveau α fixé, t est fixé, la règle de décision est fixée, et **on n'a plus aucun moyen de contrôler le deuxième type d'erreur**, à savoir

$$\text{erreur de deuxième espèce} = P_{(H_1)}(\text{accepter } (H_0)).$$

Dans l'exemple précédent, cette erreur est donnée par la fonction

$$p \in]0; 0.1[\mapsto P_p \left[\bar{X}_n \geq 0.1 - 1.645 \sqrt{\frac{(0.1)(0.9)}{n}} \right].$$

Si on admet que cette fonction est continue (ce qui sera toujours le cas dans le cadre de ce cours), et que l'on fait $p \rightarrow 10\%$, la valeur de cette erreur tend vers

$$\begin{aligned} &P_{p=10\%} \left[\bar{X}_n \geq 0.1 - 1.645 \sqrt{\frac{(0.1)(0.9)}{n}} \right] \\ &= 1 - P_{p=10\%} \left[\bar{X}_n < 0.1 - 1.645 \sqrt{\frac{(0.1)(0.9)}{n}} \right] = 1 - 5\% = 95\%, \end{aligned}$$

par définition de t . Ainsi, l'erreur de deuxième espèce peut être élevée suivant les valeurs du paramètre. **Il ne faut pas s'en étonner !** Plus p est proche de 10% (par valeurs inférieures), et plus il est difficile de voir sur un échantillon si p est plus grand ou plus petit que 10%.

La méconnaissance de l'erreur de deuxième espèce a la conséquence suivante :

la conclusion d'un test n'a valeur de preuve que lorsque cette conclusion est le rejet de (H_0) .

Ainsi il faut plutôt voir l'acceptation de (H_0) comme un non-rejet de (H_0) . La démarche est ici très empirique : sur l'expérience que j'ai faite, rien ne permet de dire que (H_0) n'est pas vraie.

La construction d'un test suit alors les étapes suivantes :

1. **Choix du niveau α** (petit - en général = 5%).
2. **Choix de l'hypothèse privilégiée (H_0)** , qui doit être choisie en fonction de l'erreur que l'on veut contrôler : $\alpha = P_{(H_0)}(\text{rejeter } H_0)$.
3. **Choix d'un estimateur $\hat{\theta}$** du paramètre θ sur lequel porte le test, **dont on connaisse la loi sous (H_0)** (\bar{X}_n dans l'exemple du modèle des ampoules).
4. **Construction d'une région de rejet de (H_0) de niveau α** . Celle-ci sera souvent construite sur l'estimateur $\hat{\theta}$, et sera du type $\mathcal{R} = \{\hat{\theta} \in R\}$ ($\{\bar{X}_n < t\}$ dans l'exemple précédent).

La règle de décision correspondante est :

- si sur l'échantillon $\hat{\theta}(\omega) \in R$, on rejette (H_0) ;
- si sur l'échantillon $\hat{\theta}(\omega) \notin R$, on ne rejette pas (H_0) .

L'ensemble R est à choisir de telle sorte que $\alpha = P_{(H_0)}[\hat{\theta} \in R]$. La construction effective du test nécessite donc bien la connaissance de la loi de $\hat{\theta}$ sous (H_0) .

5. **Conclusion au vu de l'échantillon de données.**
6. **Si la conclusion est l'acceptation de (H_0) , étude éventuelle de l'erreur de deuxième espèce.**

Remarque On peut voir sur l'exemple précédent que lors du calcul de t , seule la valeur $p = 10\%$ est intervenue. On aurait donc construit le même test si on avait testé

$$(H_0) : p = 10\% \quad \text{contre} \quad (H_1) : p < 10\% .$$

Ce sera le cas dans toutes les situations rencontrées dans ce cours. Cela traduit le fait intuitif suivant : plus p est loin de 10% (par valeurs supérieures), plus il est facile de voir sur un échantillon de $\mathcal{B}(p)$ que $p \geq 10\%$. Le cas le plus délicat à traiter correspond au cas $p = 10\%$. Aussi, dans la suite, on prendra toujours des hypothèses (H_0) sous forme d'égalité, ce qui évite le passage (superflu) par le "sup".

En pratique, au lieu de fixer un niveau α , et d'en déduire le $t(\alpha)$ qui lui est associé par la résolution de (9.1), il est plus commode de calculer la **P-valeur** du test. Dans l'exemple précédent où l'on teste

$$(H_0) : p = 10\% (p \geq 10\%) \quad \text{contre} \quad (H_1) : p < 10\% ,$$

la région de rejet est du type $\{\bar{X}_n < t\}$. Supposons que sur notre échantillon, on ait observé une valeur $\bar{X}_n(\omega) = \bar{x}_n$ ($= 0.09$ pour fixer les idées), la P -valeur du test est donnée par

$$\alpha_{\text{obs}} = P_{(H_0)}(\bar{X}_n < \bar{x}_n) = P_{p=0.1}(\bar{X}_n < 0.09).$$

Une première observation est donc que la P -valeur permet de décider si l'on rejette ou non (H_0), quel que soit le niveau α choisi, par la règle suivante

1. Si $\alpha > \alpha_{\text{obs}}$, on rejette (H_0).
2. Si $\alpha < \alpha_{\text{obs}}$, on ne rejette pas (H_0).

En fait on peut montrer (exercice!) que sous (H_0), α_{obs} est la réalisation d'une variable aléatoire de loi uniforme sur $[0, 1]$, alors que sous (H_1), par construction de la zone de rejet, α_{obs} a tendance à prendre des valeurs proches de 0. Ainsi si le niveau α du test n'est pas imposé de l'extérieur (par ex. par des normes), il peut être plus intéressant d'évaluer le risque pris en rejetant (H_0) directement à partir de la P -valeur (plutôt que de s'imposer une règle stricte de décision en fixant à l'avance le niveau du test α). À moins bien sûr d'avoir peur de se laisser influencer par une trop forte envie de prouver (H_1), et d'être ainsi amené à rejeter (H_0) même avec un α_{obs} élevé.

Pour y voir plus clair, terminons le calcul de la P -valeur sur notre exemple. On obtient

$$\begin{aligned} P_{\text{val}} &= P_{p=0.1}(\bar{X}_n < 0.09) = P_{p=0.1} \left[\frac{\bar{X}_n - 0.1}{\sqrt{(0.1)(0.9)/1000}} < \frac{-0.01}{\sqrt{(0.1)(0.9)/1000}} \right] \\ &= P[Z < -1.05], \text{ avec } Z \sim \mathcal{N}(0, 1), \\ &= P[Z > 1.05] = 14,68\%, \end{aligned}$$

d'après les tables de la loi normale. Ainsi si on accepte de se tromper dans 15% des cas, on peut conclure que $p < 1/10$. Dans le cas contraire, rien ne nous permet de conclure que $p < 1/10$.

9.2 Échantillons gaussiens.

On suppose maintenant que (X_1, \dots, X_n) est un n -échantillon de loi $\mathcal{N}(\mu, \sigma^2)$.

9.2.1 Test sur μ avec σ connu.

Dans l'exemple 7.2, supposons que le chimiste veuille savoir si la concentration μ en acétone du produit considéré est supérieure à une valeur donnée μ_0 ou pas. Supposons par ailleurs qu'il préfère se tromper en affirmant que $\mu \geq \mu_0$ plutôt qu'en affirmant que $\mu < \mu_0$. L'erreur qu'il veut contrôler est donc

$$P_{\mu \geq \mu_0} [\text{conclure que } \mu < \mu_0] = P_{H_0} [\text{rejeter } H_0].$$

Il va donc choisir de tester

$$(H_0) : \mu = \mu_0 \text{ (en fait } \mu \geq \mu_0) \text{ contre } (H_1) : \mu < \mu_0.$$

Pour se décider entre ces deux hypothèses, il fait une série de n dosages X_1, \dots, X_n , qu'il suppose être un n -échantillon de loi $\mathcal{N}(\mu, \sigma^2)$, où σ est connu.

Sa règle de décision est basée sur l'estimateur \bar{X}_n de μ , qui est de loi connue $\mathcal{N}(\mu_0, \frac{\sigma^2}{n})$, sous l'hypothèse (H_0) . Si \bar{X}_n est "petit", alors il peut raisonnablement penser que $\mu < \mu_0$. Ainsi, la règle de décision est du type :

1. Si $\bar{X}_n \geq t$, il décide $\mu \geq \mu_0$ (acceptation de (H_0)).
2. Si $\bar{X}_n < t$, il décide $\mu < \mu_0$ (rejet de (H_0)).

Supposons que sur sa série de n dosages, il a observé une valeur \bar{x}_n de \bar{X}_n . La P-valeur associée à sa règle de décision est donc

$$P_{val} = P_{(H_0)} [\bar{X}_n < \bar{x}_n] = P_{\mu=\mu_0} \left[\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} < \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} \right].$$

Quand $\mu = \mu_0$, $\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$, et on a donc

$$P_{val} = P \left[Z < \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} \right], Z \sim \mathcal{N}(0, 1).$$

Connaissant \bar{x}_n (moyenne empirique sur son échantillon), μ_0 (valeur à tester), σ et n , il lit P_{val} dans la table de la loi normale. Sa conclusion est la suivante :

1. Pour un niveau $\alpha > P_{val}$, rejet de (H_0) .
2. Pour un niveau $\alpha < P_{val}$, non rejet de (H_0) .

9.2.2 Test sur μ avec σ inconnu.

Dans l'exemple précédent, supposons que le chimiste ne connaît pas la précision de la méthode de dosage qu'il a utilisée. Sa série de n dosages (X_1, \dots, X_n) est donc maintenant un n -échantillon de loi $\mathcal{N}(\mu, \sigma^2)$, où μ et σ sont inconnus.

Supposons qu'il veut à nouveau tester

$$(H_0) : \mu = \mu_0 \text{ (en fait } \mu \geq \mu_0) \text{ contre } (H_1) : \mu < \mu_0.$$

S'il refait le raisonnement précédent, il se trouve incapable de calculer la P-valeur

$$P_{val} = P \left[Z < \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} \right],$$

puisqu'il ne connaît pas σ .

Il va donc utiliser la variable $T_n = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n-1}}$, qui sous l'hypothèse $(H_0) : \mu = \mu_0$, suit la loi de Student à $n - 1$ degrés de liberté. Si T_n est "très négatif", il peut raisonnablement penser que $\mu < \mu_0$. Ainsi, sa règle de décision est du type :

1. Si $\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n-1}} \geq t$, il décide $\mu = \mu_0$ (non rejet de (H_0)).
2. Si $\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n-1}} < t$, il décide $\mu < \mu_0$ (rejet de (H_0)).

Supposons que sur sa série de n dosages, il a observé une valeur \bar{x}_n de \bar{X}_n , et une valeur s_n de S_n . La P-valeur associée à sa règle de décision est donc

$$P_{val} = P_{(H_0)} \left[\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n-1}} < \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n-1}} \right] = P_{\mu=\mu_0} \left[\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n-1}} < \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n-1}} \right].$$

Quand $\mu = \mu_0$, $\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n-1}} \sim \mathcal{S}(n-1)$, et on a donc

$$P_{val} = P \left[Z < \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n-1}} \right], \quad Z \sim \mathcal{S}(n-1).$$

Il lit P_{val} dans les tables de lois de Student. Sa conclusion est la suivante :

1. Pour un niveau $\alpha > P_{val}$, rejet de (H_0) .
2. Pour un niveau $\alpha < P_{val}$, non rejet de (H_0) .

9.2.3 Test sur σ .

Supposons toujours que le chimiste ne connaît pas la précision σ de la méthode de dosage, et qu'il veut savoir si cette précision est meilleure qu'une précision donnée σ_0 , ou pas. On peut supposer qu'il préfère se tromper en concluant que $\sigma \geq \sigma_0$ (auquel cas ses résultats sont plus précis que ce qu'il a annoncé), plutôt qu'en concluant que $\sigma < \sigma_0$. Il veut donc contrôler

$$P_{\sigma \geq \sigma_0} [\text{conclure } \sigma < \sigma_0].$$

Ainsi, il choisit de tester

$$(H_0) : \sigma = \sigma_0 \text{ (en fait } \sigma \geq \sigma_0) \text{ contre } (H_1) : \sigma < \sigma_0.$$

Sa règle de décision est basée sur l'estimateur S_n^2 de σ^2 , qui sous l'hypothèse (H_0) est de loi connue ($nS_n^2/\sigma_0^2 \sim \chi^2(n-1)$ sous (H_0)). Si S_n^2/σ_0^2 est "petit", il peut raisonnablement penser que $\sigma < \sigma_0$. Ainsi, la règle de décision est du type :

1. Si $\frac{nS_n^2}{\sigma_0^2} < t$, il décide $\sigma < \sigma_0$ (rejet de (H_0)).
2. Si $\frac{nS_n^2}{\sigma_0^2} \geq t$, il décide $\sigma \geq \sigma_0$ (non rejet de (H_0)).

Supposons que sur sa série de n dosages, il observe une valeur s_n^2 de S_n^2 . La P-valeur associée à sa règle de décision est donc

$$P_{val} = P_{\sigma=\sigma_0} \left[\frac{nS_n^2}{\sigma_0^2} < \frac{ns_n^2}{\sigma_0^2} \right].$$

Quand $\sigma = \sigma_0$, $\frac{nS_n^2}{\sigma_0^2} \sim \chi^2(n-1)$, et on a donc

$$P_{val} = P \left[Z < \frac{ns_n^2}{\sigma_0^2} \right], \quad Z \sim \chi^2(n-1).$$

Il lit P_{val} dans les tables de lois du χ^2 . Sa conclusion est la suivante :

1. Pour un niveau $\alpha > P_{val}$, rejet de (H_0) .
2. Pour un niveau $\alpha < P_{val}$, non rejet de (H_0) .

Chapitre 10

Tests du χ^2

Les tests du χ^2 font partie des tests les plus utilisés en statistique. La raison en est qu'ils peuvent s'appliquer dans beaucoup de situations (paramétriques, ou **non paramétriques**). La plupart de ces tests peuvent être utilisés pour vérifier qu'un modèle choisi est en adéquation avec la réalité, et peuvent donc aider le statisticien dans le choix de son modèle.

10.1 Test du χ^2 d'ajustement.

La problématique des tests d'ajustement est la suivante. Étant donné un n -échantillon d'une loi μ **inconnue** sur un espace E (connu), on désire vérifier si $\mu = \mu_0$, où μ_0 est une **probabilité** sur E **donnée**. Par exemple, on veut savoir si le nombre quotidien d'accidents de voitures dans une ville suit une loi de Poisson de paramètre 10. Dans cet exemple, $E = \mathbb{N}$, et μ_0 est la loi $\mathcal{P}(10)$. Dans cette situation, le paramètre du modèle est l'ensemble des probabilités sur E .

De façon plus générale, on veut savoir si μ appartient à un certain ensemble de lois, ou pas. Par exemple, on veut savoir si le nombre quotidien d'accidents de voitures dans une ville suit une loi de Poisson.

Le test du χ^2 d'ajustement permet d'apporter une réponse à cette question.

10.1.1 Test du χ^2 d'ajustement à une loi sur un espace fini.

C'est la version la plus simple du test du χ^2 . L'observation est une réalisation d'un n -échantillon (X_1, \dots, X_n) de la loi $p = \sum_{j=1}^d p_j \delta_{\xi_j}$, où les ξ_j sont connus, et les p_j sont les paramètres du modèle.

Soit $\pi = \sum_{j=1}^d \pi_j \delta_{\xi_j}$ (π_j connus). On veut tester $(H_0) : "p = \pi"$ contre $(H_1) : "p \neq \pi"$.

Exemple 10.1.1. bauer Partant de races pures, Bauer a croisé des mufliers ivoires avec des mufliers rouges. A la deuxième génération, après autofécondation des plantes de la première, il a obtenu : 22 mufliers rouges, 52 pâles et 23 ivoires. Si la couleur des fleurs est gérée par un couple d'allèles, la répartition théorique est : *Rouge* $\rightarrow 1/4$, *Pâle* $\rightarrow 1/2$, *Ivoire* $\rightarrow 1/4$.

Que conclure ?

Dans cet exemple, les observations sont $(X_1(\omega), \dots, X_{97}(\omega))$, où $X_i(\omega)$ est la couleur de la i -ème fleur. Les X_i peuvent prendre trois valeurs $\{R, P, I\}$ (les " ξ_j ") avec probabilité p_R, p_P, p_I (les " p_j "). Il s'agit de voir si $p_R = 1/4, p_P = 1/2, p_I = 1/4$.

Si n est grand, les fréquences empiriques N_j/n où $N_j = \sum_{i=1}^n \mathbb{1}_{X_i=\xi_j}$ ($j \in \{1, \dots, d\}$) est le nombre de fois où on rencontre ξ_j dans l'échantillon, sont proches de $P(X_1 = \xi_j) = p_j$ par la loi des grands nombres. Karl Pearson a donc introduit la statistique

$$T_n = \sum_{j=1}^d \frac{1}{n\pi_j} (N_j - n\pi_j)^2$$

qui devrait être proche de 0 dans l'hypothèse (H_0) .

L'idée est alors de construire un test dont la région de rejet serait du type $\{T_n \geq t\}$. A cette fin, on a besoin de connaître la loi de T_n sous (H_0) .

Théorème 10.2 (admis) *On suppose que $\forall j \in \{1, \dots, d\}, \pi_j > 0$. Si les variables $(X_i)_{i=1, \dots, n}$ sont i.i.d. de loi π ,*

$$(loi) \quad T_n \xrightarrow[n \rightarrow \infty]{} Z, \quad \text{où } Z \sim \chi_{d-1}^2.$$

Région de rejet du test du χ^2 d'ajustement à π . Soit α fixé, et t_α défini par

$$P[X > t_\alpha] = \alpha, \quad \text{où } X \sim \chi_{d-1}^2. \quad \mathcal{R} = \left\{ \sum_{j=1}^d \frac{(N_j - n\pi_j)^2}{n\pi_j} > t_\alpha \right\}$$

est une zone de rejet de (H_0) : " $p = \pi$ " contre (H_1) : " $p \neq \pi$ " de niveau α , si **n est assez grand**. En pratique, on impose $n \geq 30, n\pi_j \geq 5$, pour tout $j \in \{1, \dots, d\}$. de façon traditionnelle, les variables N_j sont appelés les **effectifs observés**, les nombres $n\pi_j$ sont appelés les **effectifs théoriques**.

Région de confiance pour π . Le théorème 10.2 peut s'interpréter en disant que la variable T_n est un pivot asymptotique pour l'estimation de la loi des X_j . On en déduit donc des régions de confiance pour cette loi. Soit t_α défini comme précédemment. On a alors, si n est assez grand,

$$\forall \pi, P_\pi \left[\sum_{i=1}^d \frac{(N_i - n\pi(i))^2}{n\pi(i)} \leq t_\alpha \right] \simeq 1 - \alpha.$$

10.1.2 Test du χ^2 d'ajustement à une loi quelconque

Le test du χ^2 d'ajustement peut également être utilisé pour tester l'ajustement à une loi à densité Q sur un ensemble E quelconque connu, i.e. pour tester

$$(H_0) : (X_1, \dots, X_n) \text{ est un } n - \text{échantillon de loi } Q, \\ \text{contre } (H_1) : (X_1, \dots, X_n) \text{ n'est pas un } n - \text{échantillon de loi } Q.$$

Q est ici une loi **connue** (par exemple, la loi $\mathcal{N}(0, 1)$). Pour cela, on commence par choisir un entier d et une application $\phi : E \rightarrow \{1, \dots, d\}$, et on applique ensuite le test du χ^2 à l'échantillon $(Y_1, \dots, Y_n) = (\phi(X_1), \dots, \phi(X_n))$, qui est un n -échantillon d'une loi sur l'ensemble fini $\{1, \dots, d\}$.

Se donner une telle application ϕ , revient à se donner une partition V_1, \dots, V_d de E et à poser $\phi(x) = j, \forall x \in V_j$. Sous (H_0) , on a alors

$$\begin{aligned} \pi_j &= P(Y_1 = j) = P(X_1 \in V_j) = Q(V_j), \\ N_j &= \sum_{i=1}^n \mathbb{1}_{Y_i=j} = \sum_{i=1}^n \mathbb{1}_{X_i \in V_j} = \text{nombre de } X_i \text{ qui tombent dans } V_j. \end{aligned}$$

Au niveau α , la région de rejet du test du χ^2 d'ajustement à la loi Q est alors

$$\mathcal{R} = \left\{ \sum_{i=1}^d \frac{(N_j - nQ(V_j))^2}{nQ(V_j)} \geq t_\alpha \right\}, \text{ où } t_\alpha \text{ est tel que } P[Z \geq t_\alpha] = \alpha, Z \sim \chi_{d-1}^2.$$

Notez que ce test dépend de la partition choisie. Il y a donc autant de tests du χ^2 dans cette situation, que de choix de partitions. La seule contrainte sur le choix de la partition est que $\forall j, nQ(V_j) \geq 5$. Les conclusions du test peuvent être très différentes suivant la choix de la partition. Par exemple, nous présentons dans la figure ??, les Pvaleurs des tests du χ^2 d'ajustement à la loi $\mathcal{N}(0, 1)$ pour différentes partitions de \mathbb{R} (au nombre de 1000), pour **une même réalisation** $(X_1(\omega), \dots, X_{1000}(\omega))$ (générée avec le générateur de nombres aléatoires $\mathcal{N}(0, 1)$ de matlab).

10.1.3 Test du χ^2 d'ajustement à une famille de loi.

Le test du χ^2 peut être généralisé au cas où on se permet d'estimer un certain nombre de paramètres de la loi. Commençons par présenter la situation dans le cas où les observations ne peuvent prendre qu'un nombre fini de valeurs connues. Dans ce cas, on suppose que l'observation (X_1, \dots, X_n) est un n -échantillon de loi $\mu(\theta) = \sum_{j=1}^d p_j(\theta) \delta_{\xi_j}$, les ξ_j étant connus et $\theta \in \Theta \subset \mathbb{R}^k$ étant un paramètre inconnu. On veut alors tester

$$\begin{aligned} (H_0) : & (X_1, \dots, X_n) \text{ est un } n - \text{échantillon de loi } \in \{\mu(\theta), \theta \in \Theta_0\} \\ \text{contre } (H_1) : & (X_1, \dots, X_n) \text{ est un } n - \text{échantillon de loi } \in \{\mu(\theta), \theta \in \Theta_1\} \end{aligned}$$

Pour cela, on introduit

$$T_n(\theta) = \sum_{j=1}^d \frac{1}{np_j(\theta)} (N_j - np_j(\theta))^2$$

$T_n(\theta)$ n'est pas une statistique puisqu'elle dépend du paramètre θ . On estime alors θ par l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ et on choisit comme région de rejet du test

$$R(X) = \left\{ T_n(\hat{\theta}_n) \geq t \right\}$$

Pour déterminer t , on a besoin de connaître la loi de $T_n(\hat{\theta}_n)$ sous (H_0) . On admettra le résultat suivant :

Théorème 10.3 *Supposons que*

- Θ_0 est un ouvert de \mathbb{R}^k ;
- $\forall j \in \{1, \dots, d\}$, $p_j : \Theta_0 \rightarrow [0; 1]$ est de classe \mathcal{C}^2 ;
- $\forall \theta \in \Theta_0$, $\left(\frac{\partial p_j}{\partial \theta_l} \right)_{\substack{j=1 \dots d \\ l=1 \dots k}}$ est de rang k (ce qui impose $k \leq d$) ;
- l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de $\theta \in \Theta_0$ existe pour tout n , et vérifie pour tout $j \in \{1, \dots, d\}$, $\forall n$, $p_j(\hat{\theta}_n) > 0$.

Alors, sous (H_0) , $T_n(\hat{\theta}_n) \xrightarrow{(loi)} Z$, où $Z \sim \chi_{d-k-1}^2$.

La formule presque incantatoire à retenir, est que le degré du χ^2 est égal au

nombre de classes - 1 - nombre de paramètres estimés ($d - 1 - k$).

Ainsi, la région de rejet du test du χ^2 de niveau α de

(H_0) : (X_1, \dots, X_n) est un n - échantillon de loi $\in \{\mu(\theta), \theta \in \Theta_0\}$
contre (H_1) : (X_1, \dots, X_n) est un n - échantillon de loi $\in \{\mu(\theta), \theta \in \Theta_1\}$

est donnée par

$$\mathcal{R} = \left\{ T_n(\hat{\theta}_n) \geq t_\alpha \right\}, \text{ où } t_\alpha \text{ est solution de } P[Z \geq t_\alpha] = \alpha, Z \sim \chi_{d-k-1}^2.$$

Là encore, on peut généraliser à des lois qui ne sont plus nécessairement à support fini, en projetant les données sur une partition.

Ce test peut s'utiliser pour tester l'ajustement des données à une famille de lois. Par exemple, le chimiste peut se demander si la série de dosages $(X_1(\omega), \dots, X_n(\omega))$ qu'il a obtenue, est bien la réalisation d'un échantillon gaussien. Il va donc tester

$$(H_0) : \exists (m, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ \text{ tel que } X_i \sim \mathcal{N}(m, \sigma^2) \\ \text{contre } (H_1) : \forall (m, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+, X_i \text{ ne suit pas } \mathcal{N}(m, \sigma^2)$$

On a déjà vu que dans ce contexte (cf exemple 7.3.??), l'estimateur du maximum de vraisemblance de (m, σ^2) est (\bar{X}_n, S_n^2) , où $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ et $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. Pour effectuer son test, le chimiste va commencer par choisir une partition V_1, \dots, V_d de \mathbb{R} . Sous (H_0) ,

$$P_{(H_0)} [X_1 \in V_j] = \int_{V_j} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-m)^2}{2\sigma^2}\right) dy \triangleq p_j(m, \sigma^2).$$

Si on désigne par N_i le nombre de $(X_k; k = 1, \dots, n)$ qui tombent dans V_i ($N_i \triangleq \sum_{k=1}^n \mathbb{1}_{X_k=i}$), la statistique de test que le chimiste va utiliser est donc :

$$T_n = \sum_{j=1}^d \frac{(N_j - np_j(\bar{X}_n, S_n^2))^2}{np_j(\bar{X}_n, S_n^2)}.$$

Sous (H_0) , T_n tend en loi vers une variable du χ_{d-2-1}^2 quand la taille n de l'échantillon est suffisamment grande. La région de rejet d'un test de niveau α est donc $\mathcal{R} = \{T_n \geq t_\alpha\}$, où t_α est solution de $P[Z \geq t_\alpha] = \alpha$, avec $Z \sim \chi_{d-3}^2$.

10.2 Test du χ^2 d'indépendance.

Le théorème 10.3 peut être utilisé pour tester l'indépendance entre deux échantillons de données. On est ici dans une situation où l'observation est constituée de deux n-échantillons (X_1, \dots, X_n) et (Y_1, \dots, Y_n) , et il s'agit de savoir si ces deux échantillons sont indépendants ou pas. Il revient au même de dire que l'observation est constituée d'un n-échantillon d'un couple de variables (X, Y) à valeurs dans $E \times F$. On veut tester (H_0) : “ X et Y sont indépendantes ” contre (H_1) : “ X et Y ne sont pas indépendantes ”. Pour cela, on choisit des partitions de E et F , $E = \cup_{i=1}^k E_i$, $F = \cup_{j=1}^m F_j$, et on désigne par N_{ij} le nombre de points (X_l, Y_l) qui tombent dans $E_i \times F_j$:

$$N_{ij} = \sum_{l=1}^n \mathbb{1}_{X_l \in E_i} \mathbb{1}_{Y_l \in F_j}.$$

Si on connaissait les lois marginales du couple (X, Y) , i.e. les lois de la variable X , et celle de la variable Y , on pourrait utiliser comme statistique de test

$$T = \sum_{i=1}^k \sum_{j=1}^m \frac{(N_{ij} - np_i q_j)^2}{np_i q_j},$$

où on a noté $p_i = P[X \in E_i]$ et $q_j = P[Y \in F_j]$. En effet, sous (H_0) , T tendrait en loi vers un χ_{km-1}^2 . Ici, on ne connaît pas les lois de X et de Y . On les estime donc par leurs estimateurs du maximum de vraisemblance sous (H_0) , i.e on estime p_i par $\hat{p}_i = N_{i\bullet} = \frac{1}{n} \sum_{j=1}^m N_{ij}$ (la proportion de points (X_k, Y_k) dont la coordonnée X vaut i), et q_j par $\hat{q}_j = N_{\bullet j} = \frac{1}{n} \sum_{i=1}^k N_{ij}$ (la proportion de points (X_k, Y_k) dont la coordonnée Y vaut j). On utilise alors la statistique

$$S = \sum_{i=1}^k \sum_{j=1}^m \frac{(N_{ij} - n\hat{p}_i \hat{q}_j)^2}{n\hat{p}_i \hat{q}_j}.$$

On a ainsi estimé $(k-1) + (m-1)$ paramètres (les “-1” viennent des deux relations $\sum_i p_i = \sum_j q_j = 1$). On en déduit que S tend en loi vers un χ^2 de degré $km - 1 - (k-1) - (m-1) = (k-1)(m-1)$. Au niveau α , la région de rejet du test est donc $\mathcal{R} = \{S \geq t_\alpha\}$, où t_α est déterminé par l'équation $P[Z \geq t_\alpha] = \alpha$, avec $Z \sim \chi_{(k-1)(m-1)}^2$.