10 Probabilités & Estimation

I– Suites de variables aléatoires discrètes finies

1 – Indépendance d'une famille de variables aléatoires

Définition 10.1 – Soient $n \geqslant 2$ un entier et $X_1, X_2, \dots, X_n$ des variables aléatoires définies sur Ω.
On dit que les variables aléatoires $X_1, X_2, \dots, X_n$ sont **mutuellement indépendantes** lorsque pour tout $(x_1, x_2, \dots, x_n) \in X_1(\Omega) \times X_2(\Omega) \times \cdots \times X_n(\Omega)$,

$$P\big([X_1 = x_1] \cap [X_2 = x_2] \cap \cdots \cap [X_n = x_n]\big) = P(X_1 = x_1) \times P(X_2 = x_2) \times \cdots \times P(X_n = x_n).$$

Définition 10.2 – Soit $(X_n)_{n\geqslant 1}$ une suite de variables aléatoires définies sur Ω. On dit que la suite $(X_n)_{n\geqslant 1}$ est une suite de variables aléatoires mutuellement indépendantes si et seulement si pour tout $m \in \mathbb{N}^*$, les variables aléatoires $X_1, X_2, \dots, X_m$ sont mutuellement indépendantes.

Exemple 10.3 – On lance une pièce équilibrée jusqu'à obtenir PILE. Pour tout $n \in \mathbb{N}^*$, on note X_n la variable aléatoire égale à 1 si on obtient PILE au n-ième lancer et 0 sinon.
Alors la suite $(X_n)_{n\geqslant 1}$ est une suite de variables aléatoires mutuellement indépendantes.

2 – Espérance et variance d'une famille de variables aléatoires

Théorème 10.4

Soient $n \geqslant 2$ un entier et $X_1, X_2, \dots, X_n$ des variables aléatoires définies sur Ω. Alors

$$E\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} E(X_k).$$

Autrement dit,

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n).$$

Exemple 10.5 – On considère une suite $(X_k)_{k\in\mathbb{N}^*}$ de variables aléatoires mutuellement indépendantes, dont la loi (commune) est donnée pour tout $k \in \mathbb{N}^*$ par

$$P(X_k = 1) = \frac{1}{3} \quad \text{et} \quad P(X_k = 2) = \frac{2}{3}.$$

Pour tout $n \in \mathbb{N}^*$, on pose $S_n = \sum_{k=1}^{n} X_k$. Calculer $E(S_n)$.

Théorème 10.6

Soient $n \geqslant 2$ un entier et $X_1, X_2, \ldots, X_n$ des variables aléatoires **mutuellement indépendantes** définies sur Ω. Alors

$$V\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} V(X_k).$$

Autrement dit,

$$V(X_1 + X_2 + \cdots + X_n) = V(X_1) + V(X_2) + \cdots + V(X_n).$$

Exemple 10.7 – On reprend l'exemple précédent. Calculer $V(S_n)$.

II – Inégalités classiques en théorie des probabilités

1 – Inégalité de Markov

Proposition 10.8 – Inégalité de Markov

Soit X une variable aléatoire **positive** (discrète ou à densité) admettant une espérance.
Alors pour tout réel a strictement positif,

$$P(X \geqslant a) \leqslant \frac{E(X)}{a}.$$

Remarque 10.9 – L'inégalité stricte est aussi vérifiée :

$$P(X > a) \leqslant \frac{E(X)}{a}.$$

Corollaire 10.10

Soit X une variable aléatoire (discrète ou à densité). On suppose que X^2 admet une espérance.
Alors pour tout réel a strictement positif,

$$P\big(|X| \geqslant a\big) \leqslant \frac{E\big(X^2\big)}{a^2}.$$

2 – Inégalité de Bienaymé-Tchebychev

Proposition 10.11

Soit X une variable aléatoire (discrète ou à densité). On suppose que X^2 admet une espérance.
Alors pour tout réel ε strictement positif,

$$P\big(|X - E(X)| \geqslant \varepsilon\big) \leqslant \frac{V(X)}{\varepsilon^2}.$$

Remarque 10.12 – Souvent, on reconnaît qu'il faut se servir de l'inégalité de Bienaymé-Tchebychev grâce aux valeurs absolues présentes dans la probabilité.

3 – Loi faible des grands nombres

Théorème 10.13 – Loi faible des grands nombres

Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires mutuellement indépendantes, ayant chacune la même espérance m et la même variance σ^2. On pose $\overline{X_n} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$. Alors pour tout réel ε strictement positif,

$$\lim_{n\to+\infty} P\Big(\Big|\overline{X_n} - m\Big| \geqslant \varepsilon\Big) = 0.$$

III – Estimation

Les statisticiens connaissent, en général, le type de loi qui décrit tel ou tel phénomène, par l'observation. Mais souvent ils ne connaissent pas tous les paramètres de la dite loi. Ils doivent donc les estimer : c'est l'objectif de ce que l'on appelle la statistique inférentielle.
On considère une variable aléatoire X, dont le type de loi est connu et dépend d'un paramètre réel θ inconnu (ce peut être le paramètre λ d'une variable exponentielle, l'étendue $b-a$ d'une variable uniforme sur $[a,b]$, le paramètre p d'une variable de Bernoulli, l'espérance m d'une loi normale, etc).
L'objectif est de donner une *estimation* de la vraie valeur du paramètre θ.

Il y a deux types d'estimation : l'estimation ponctuelle et l'estimation par intervalle de confiance.

1 – Échantillons et estimateurs

Exemple 10.14 – On suppose que la durée de vie (en heures) des ampoules produites par l'entreprise Lumilux suit une loi exponentielle X dont le paramètre λ est inconnu.

1. Rappeler la formule donnant l'espérance de X en fonction de λ.

2. Des tests ont été effectués sur 10 ampoules. On obtient les durées de vie suivantes :

62.1	75.4	73.1	81	68.7	73.6	64.2	78.5	74.4	63

   À l'aide de la série statistique ci-dessus, donner une estimation du paramètre λ.

3. Peut-on affirmer que λ est égal à la valeur précédente ? Si non, comment pourrait-on tenter d'améliorer la qualité de l'estimation ?

Définition 10.15 – Soient X une variable aléatoire (discrète ou à densité) et $n \in \mathbb{N}^*$ un entier.
On appelle n-**échantillon** de X tout n-uplet $(X_1, \dots, X_n)$ de variables aléatoires indépendantes et de même loi que X.

Définition 10.16 – Soient X une variable aléatoire (discrète ou à densité) et $n \in \mathbb{N}^*$ un entier.
On appelle **réalisation de l'échantillon** $(X_1, \dots, X_n)$ (ou **échantillon observé**) tout n-uplet $(x_1, \dots, x_n)$ de valeurs prises par $(X_1, \dots, X_n)$ (x_1 est la valeur prise par X_1, x_2 est la valeur prise par X_2, …, x_n est la valeur prise par X_n).

Exemple 10.17 – Dans l'exemple précédent, l'échantillon observé est

Bien évidemment, l'échantillon observé est *aléatoire*. On aurait par exemple pu, en effectuant les tests avec un autre jeu de 10 ampoules, obtenir

74.1	82.5	68.5	70.3	84	77.2	69.6	73.8	76.3	68.7

Définition 10.18 – Soient θ un réel, X une variable aléatoire dont la loi dépend d'un paramètre θ, $n \in \mathbb{N}^*$ un entier et $(X_1, \dots, X_n)$ un échantillon de X.
On appelle **estimateur** de θ toute variable aléatoire T_n de la forme $\varphi(X_1, \dots, X_n)$.

Exemple 10.19 – Dans l'exemple précédent, le paramètre que l'on cherche à estimer est le paramètre λ de la loi exponentielle suivie par X. L'estimateur considéré pour l'espérance est

$$T_n = \frac{1}{n}\sum_{k=1}^{n} X_k = \frac{1}{n}(X_1 + X_2 + \cdots + X_n).$$

Cet estimateur m'a permis d'estimer **ponctuellement**.

2 – Biais et risque quadratique d'un estimateur

Définition 10.20 – Soient θ un réel, X une variable aléatoire dont la loi dépend d'un paramètre θ et T_n un estimateur de θ.

- Si T_n admet une espérance, on appelle **biais** de T_n le réel $b(T_n)$ défini par
$$b(T_n) = E(T_n) - \theta.$$
- On dit que T_n est un **estimateur sans biais** de θ si et seulement si le biais de T_n est nul, *i.e.* si et seulement si $E(T_n) = \theta$. Dans le cas contraire, on dit que T_n est un estimateur **biaisé** de θ.

Exemple 10.21 – Soient X une variable aléatoire de loi de Bernoulli de paramètre p et $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes qui suivent toutes la même loi que X.
On pose pour tout $n \in \mathbb{N}^*$, $\overline{X_n} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$. Montrer que $\overline{X_n}$ est un estimateur sans biais de p.

Définition 10.22 – Soient θ un réel, X une variable aléatoire dont la loi dépend d'un paramètre θ et T_n un estimateur de θ.
Si T_n admet une variance, on appelle **risque quadratique** de T_n le réel $r(T_n)$ défini par

$$r(T_n) = E\left((T_n - \theta)^2\right).$$

Proposition 10.23

Soient θ un réel, X une variable aléatoire dont la loi dépend d'un paramètre θ et T_n un estimateur de θ qui admet une variance. Alors

$$r(T_n) = b(T_n)^2 + V(T_n).$$

En particulier, **si T_n est un estimateur sans biais de θ**, alors

$$r(T_n) = V(T_n).$$

3 – Estimation par intervalle de confiance

Les estimations ponctuelles ne fournissent pas d'information sur la précision des estimations, c'est-à-dire qu'elles ne tiennent pas compte de l'erreur possible attribuable aux fluctuations d'échantillonnage. Or deux échantillons distincts donnent presque certainement des valeurs distinctes pour l'estimation.
Ici, il s'agit toujours d'estimer un paramètre inconnu, mais au lieu de lui attribuer une valeur unique en faisant appel à un estimateur ponctuel, on construit un intervalle aléatoire qui permet de "recouvrir" avec une certaine fiabilité, la vraie valeur du paramètre estimé.

Définition 10.24 – Soient θ un réel, X une variable aléatoire dont la loi dépend d'un paramètre θ, U_n et V_n deux estimateurs de θ et $\alpha \in [0,1]$ un réel.
On dit que $[U_n, V_n]$ est un **intervalle de confiance** de θ au niveau de confiance $1-\alpha$ si

$$P\left(U_n \leqslant \theta \leqslant V_n\right) \geqslant 1 - \alpha.$$