9 Suites de variables aléatoires

I – Suites de variables aléatoires

1 – Indépendance d'une famille de variables aléatoires

Définition 9.1 –

- Soient $n \geqslant 2$ un entier et $X_1, X_2, \dots, X_n$ des variables aléatoires. On dit que $X_1, X_2, \dots, X_n$ sont **mutuellement indépendantes** lorsque pour tout n-uplet d'intervalles $(I_1, I_2, \dots, I_n)$,
$$P(X_1 \in I_1, X_2 \in I_2, \dots, X_n \in I_n) = P(X_1 \in I_1) \times P(X_2 \in I_2) \times \cdots \times P(X_n \in I_n).$$
- Soit $(X_n)_{n\geqslant 1}$ une suite de variables aléatoires. On dit que la suite $(X_n)_{n\geqslant 1}$ est une suite de variables aléatoires mutuellement indépendantes si et seulement si pour tout entier $m \in \mathbb{N}^*$, les variables aléatoires $X_1, X_2, \dots, X_m$ sont mutuellement indépendantes.

Exemple 9.2 – 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 9.3 – Linéarité de l'espérance

Soient $n \geqslant 2$ un entier et $X_1, X_2, \dots, X_n$ des variables aléatoires. 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 9.4 – 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 $S_n = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$. Calculer $E(S_n)$.

Je commence par calculer $E(X_1)$: $E(X_1) = 1 \times \dfrac{1}{3} + 2 \times \dfrac{2}{3} = \dfrac{1+4}{3} = \dfrac{5}{3}$.

Ainsi pour tout $k \in [\![1, n]\!]$, $E(X_k) = E(X_1) = \dfrac{5}{3}$. Je passe ensuite au calcul de l'espérance de S_n : par linéarité,
$$E(S_n) = E\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n} \times \sum_{k=1}^{n} E(X_k) = \frac{1}{n} \times \sum_{k=1}^{n} \frac{5}{3} = \frac{1}{n} \times n \times \frac{5}{3} = \frac{5}{3}.$$

Théorème 9.5

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

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

Autrement dit, sous réserve d'indépendance,

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

Exemple 9.6 – On reprend l'exemple d'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 $S_n = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$. Calculer $V\big(S_n\big)$.

Je commence par calculer $V(X_1)$. J'ai déjà calculé $E(X_1) = \dfrac{5}{3}$.

Il me faut maintenant calculer $E\big(X_1^2\big)$: $E\big(X_1^2\big) = 1^2 \times \dfrac{1}{3} + 2^2 \times \dfrac{2}{3} = \dfrac{1+8}{3} = \dfrac{9}{3} = 3.$

Donc d'après la formule de König-Huygens,

$$V(X_1) = E\big(X_1^2\big) - E(X_1)^2 = 3 - \left(\frac{5}{3}\right)^2 = \frac{27-25}{9} = \frac{2}{9}.$$

Je passe ensuite au calcul de la variance de S_n : par indépendance des X_k,

$$V\big(S_n\big) = V\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2} \times \sum_{k=1}^{n} V(X_k) = \frac{1}{n^2} \times \sum_{k=1}^{n} \frac{2}{9} = \frac{1}{n^2} \times n \times \frac{2}{9} = \frac{2}{9n}.$$

II – Inégalités classiques en théorie des probabilités

1 – Inégalité de Markov

Proposition 9.7 – Inégalité de Markov

Soit X une variable aléatoire **positive** admettant une espérance. Alors

$$\forall a > 0, \quad P(X \geqslant a) \leqslant \frac{E(X)}{a}.$$

Remarque 9.8 –

- Cette inégalité se vérifie directement tant pour une variable aléatoire discrète que à densité.
- Si une variable aléatoire X^2 admet une espérance, alors en particulier

$$\forall a > 0, \quad P\big(|X| \geqslant a\big) = P\big(X^2 \geqslant a^2\big) \leqslant \frac{E\big(X^2\big)}{a^2}.$$

2 – Inégalité de Bienaymé-Tchebychev

Proposition 9.9 – Inégalité de Bienaymé-Tchebychev

Soit X une variable aléatoire, discrète ou à densité. On suppose que X admet une variance (donc que X^2 admet une espérance). Alors

$$\forall \varepsilon > 0, \quad P\big(|X - E(X)| \geqslant \varepsilon\big) \leqslant \frac{V(X)}{\varepsilon^2}.$$

Remarque 9.10 – Il s'agit de l'inégalité de Markov appliquée à la variable aléatoire $\big(X - E(X)\big)^2$.

3 – Loi faible des grands nombres

Théorème 9.11 – 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. On pose $\overline{X_n} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$.

Alors pour tout réel ε strictement positif,

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

Remarque 9.12 – On reconnaît qu'il faut se servir de la loi faible des grands nombres lorsqu'il s'agit de calculer la limite d'une probabilité. Le sens de l'inégalité peut être retourné en considérant l'événement contraire. La seule présence de valeurs absolues dans la probabilité conduit elle à penser à l'utilisation de l'inégalité de Bienaymé-Tchebychev.

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 existe deux façons de répondre à un tel problème : l'estimation ponctuelle et l'estimation par intervalle de confiance.

Méthode 9.13 – Estimer ponctuellement le paramètre d'une loi

Pour estimer ponctuellement le paramètre d'une loi lorsque l'on connaît un échantillon de réalisations de cette loi, il suffit de faire la moyenne des valeurs observées.

Exemple 9.14 – On considère un dé non équilibré que l'on lance 500 fois et on compte le nombre de fois où l'on obtient les différentes faces du dé.

Face du dé	1	2	3	4	5	6
Nombre d'apparitions	110	87	42	69	91	101

On considère la variable aléatoire X qui vaut 1 si le dé donne 6 et 0 sinon.

1. Reconnaître la loi suivie par X.
   Je reconnais en X une variable aléatoire qui suit une loi de Bernoulli de succès "obtenir 6" et de paramètre p inconnu.

2. Donner une estimation à 0.01 près du paramètre de cette loi.
   Dans l'expérience décrite ci-dessus, on observe 500 réalisations de X que je note $x_1, x_2, \dots, x_{500}$.
   Étant donnée la définition de X, je constate que lors de cette expérience, la variable aléatoire X a pris 101 fois la valeur 1 et 399 fois la valeur 0.
   Une estimation du paramètre p de cette loi est donc donnée par

$$\frac{1}{500} \times \sum_{i=1}^{500} x_i = \frac{101}{500} \approx 0.20.$$

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 une estimation ponctuelle, on construit un intervalle aléatoire qui permet de "recouvrir" avec une certaine fiabilité, la vraie valeur du paramètre estimé.

Définition 9.15 – Soient X une variable aléatoire suivant une loi de Bernoulli de paramètre p inconnu et $a \in]0,1[$. On appelle **intervalle de confiance** au niveau $1-a$, tout intervalle I tel que

$$P(p \in I) \geqslant 1-a.$$

Il s'agit d'un intervalle qui contient le paramètre p à estimer avec une probabilité minimale donnée.

Méthode 9.16 – Déterminer un intervalle de confiance
Déterminer un intervalle de confiance consiste à trouver un intervalle dont on sait qu'il contient le paramètre inconnu p d'une variable aléatoire de Bernoulli, avec une probabilité minimale donnée.
Les différentes étapes pour déterminer cet intervalle de confiance sont toujours détaillées par l'énoncé. L'idée est d'appliquer l'inégalité de Bienaymé-Tchebychev, puis de choisir une valeur pour ε assurant le seuil de probabilité demandé. On donne ci-dessous un enchaînement classique de questions menant à la détermination d'un intervalle de confiance.

Exemple 9.17 – On suppose que le paramètre p d'une loi de Bernoulli est inconnu et on cherche à l'estimer. Pour ce faire, on considère une suite de variables aléatoires $(X_k)_{k\in\mathbb{N}^*}$ indépendantes, suivant toutes la loi de Bernoulli de paramètre p.
Pour tout $n \in \mathbb{N}^*$, on pose $S_n = \dfrac{1}{n}\sum_{k=1}^{n} X_k$.

1. a) Montrer que $E(S_n) = p$.
   Comme X_1 suit une loi Bernoulli, alors $E(X_1) = p$ et pour tout $k \in [\![1,n]\!]$, $E(X_k) = p$.
   Ainsi par linéarité de l'espérance,

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

   b) Calculer la variance de S_n.
   Comme X_1 suit une loi Bernoulli, alors $V(X_1) = p(1-p)$ et $\forall k \in [\![1,n]\!]$, $V(X_k) = p(1-p)$.
   Ainsi par indépendance des X_k,

$$V(S_n) = V\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2} \times \sum_{k=1}^{n} V(X_k) = \frac{1}{n^2} \times \sum_{k=1}^{n} p(1-p) = \frac{1}{n^2} \times n \times p(1-p) = \frac{p(1-p)}{n}.$$

2. a) On admet que pour tout $p \in]0,1[$, $p(1-p) \leqslant \dfrac{1}{4}$. À l'aide de l'inégalité de Bienaymé-Tchebychev, montrer que pour tout $\varepsilon > 0$,

$$P\left(|S_n - p| \leqslant \varepsilon\right) \geqslant 1 - \frac{1}{4n\varepsilon^2}.$$

J'applique l'inégalité de Bienaymé-Tchebychev à la variable aléatoire S_n.
D'après les questions précédentes, $E(S_n) = p$ et $V(S_n) = \dfrac{p(1-p)}{n}$.
J'obtiens ainsi pour tout $\varepsilon > 0$,

$$P\left(|S_n - p| \geqslant \varepsilon\right) \leqslant \frac{\frac{p(1-p)}{n}}{\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2}.$$

Or d'après l'énoncé, $p(1-p) \leqslant \dfrac{1}{4}$. Donc

$$P\left(|S_n - p| \geqslant \varepsilon\right) \leqslant \frac{p(1-p)}{n\varepsilon^2} \leqslant \frac{1}{4n\varepsilon^2}.$$

Puis par complémentarité, $P\left(|S_n - p| \geqslant \varepsilon\right) = 1 - P\left(|S_n - p| < \varepsilon\right)$.
J'obtiens donc

$$1 - P\left(|S_n - p| < \varepsilon\right) \leqslant \frac{1}{4n\varepsilon^2},$$

i.e.

$$P\left(|S_n - p| < \varepsilon\right) \geqslant 1 - \frac{1}{4n\varepsilon^2}.$$

Et puisque $P\left(|S_n - p| < \varepsilon\right) \leqslant P\left(|S_n - p| \leqslant \varepsilon\right)$, alors j'obtiens bien le résultat demandé.

b) Montrer que $|S_n - p| \leqslant \varepsilon \iff p \in \left[S_n - \varepsilon, S_n + \varepsilon\right]$.
Je raisonne par équivalence :

$$\begin{aligned} |S_n - p| \leqslant \varepsilon &\iff -\varepsilon \leqslant S_n - p \leqslant \varepsilon \iff S_n - \varepsilon \leqslant p \leqslant S_n + \varepsilon \\ &\iff p \in \left[S_n - \varepsilon, S_n + \varepsilon\right]. \end{aligned}$$

c) En déduire que l'intervalle $\left[S_n - \sqrt{\dfrac{5}{n}}, S_n + \sqrt{\dfrac{5}{n}}\right]$ est un intervalle de confiance de p au niveau de confiance 0.95.

À l'aide de la question **2.b)**, l'inégalité de la question **2.a)** se réécrit

$$\forall \varepsilon > 0, \quad P\Big(p \in \left[S_n - \varepsilon, S_n + \varepsilon\right]\Big) \geqslant 1 - \frac{1}{4n\varepsilon^2}.$$

Je dois maintenant choisir ε de telle sorte que $1 - \dfrac{1}{4n\varepsilon^2} = 0.95$.
Je raisonne par équivalence :

$$1 - \frac{1}{4n\varepsilon^2} = 0.95 \iff \frac{1}{4n\varepsilon^2} = \frac{5}{100} \iff \frac{1}{\varepsilon^2} = \frac{n}{5} \iff \varepsilon = \sqrt{\frac{5}{n}}.$$

Ainsi en prenant $\varepsilon = \sqrt{\dfrac{5}{n}}$, j'obtiens bien

$$P\left(p \in \left[S_n - \sqrt{\frac{5}{n}}, S_n + \sqrt{\frac{5}{n}}\right]\right) \geqslant 0.95.$$

Cela signifie que l'intervalle $\left[S_n - \sqrt{\dfrac{5}{n}}, S_n + \sqrt{\dfrac{5}{n}}\right]$ est bien un intervalle de confiance de p au niveau 0.95.