# 10 Probabilités & Estimation

## I– Inégalités classiques en théorie des probabilités

### 1 – Inégalité de Markov

**Proposition 10.1 – Inégalité de Markov**

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

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

**Remarque 10.2 –** On a également

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

**Corollaire 10.3**

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

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

### 2 – Inégalité de Bienaymé-Tchebychev

**Proposition 10.4**

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

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

**Remarque 10.5 –** 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 10.6 – Loi faible des grands nombres**

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

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

# II – 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.

Considérons une variable aléatoire $X$, dont le type de loi est connu et dépend d'un paramètre réel $\theta$ inconnu (ce peut être le paramètre $\lambda$ 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 $\theta$.

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

## 1 – Échantillons et estimateurs

**Exemple 10.7 –** On suppose que la durée de vie (en heures) des ampoules produites par l'entreprise Lumilux suit une loi exponentielle $X$ dont le paramètre $\lambda$ est inconnu.

1. Rappeler la formule donnant l'espérance de $X$ en fonction de $\lambda$.
   On a $E(X) = \dfrac{1}{\lambda}$.
2. Des tests ont été effectués sur 10 ampoules. On a obtenu 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 $\lambda$.
   On sait que l'espérance correspond à la *moyenne* des valeurs prises par $X$. Or la moyenne de la série statistique ci-dessus vaut
   $$\frac{62.1+75.4+\cdots+63}{10} \approx 71.4.$$
   Ainsi on peut donc estimer que $\dfrac{1}{\lambda} \approx 71.4$. Et donc $\lambda \approx \dfrac{1}{71.4} \approx 0.014$.
3. Peut-on affirmer que $\lambda$ est égal à la valeur précédente? Si non, comment pourrait-on tenter d'améliorer la qualité de l'estimation?
   On ne peut pas affirmer que $\lambda$ est égal à la valeur précédente car il s'agit ici uniquement d'un échantillon de valeurs prises par $X$. Une autre série de 10 mesures donnerait une autre estimation (*a priori* proche) du paramètre $\lambda$. Pour tenter d'améliorer la qualité de l'estimation, il faudrait obtenir un échantillon de taille plus importante.

**Définition 10.8 –** Soit $X$ 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 10.9 –** Soit $X$ 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 10.10 –** Dans l'exemple introductif, l'échantillon observé est

$$(62.1,\ 75.4,\ 73.1,\ 81,\ 68.7,\ 73.6,\ 64.2,\ 78.5,\ 74.4,\ 63).$$

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

**Exemple 10.12 –** Dans l'exemple introductif, le paramètre que l'on cherche à estimer est le paramètre $\lambda$ de la loi exponentielle suivie par $X$. L'estimateur considéré pour l'espérance est

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

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

## 2 – Biais et risque quadratique d'un estimateur

**Définition 10.13 –** Soient $\theta$ un réel, $X$ une variable aléatoire dont la loi dépend d'un paramètre $\theta$ et $T_n$ un estimateur de $\theta$.

- 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 $\theta$ 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 $\theta$.

**Exemple 10.14 –** Soit $X$ une variable aléatoire de loi de Bernoulli de paramètre $p$. Soient $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes qui suivent toutes la loi de $X$.
On pose pour tout $n$ de $\mathbb{N}^*$, $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Montrer que $\overline{X_n}$ est un estimateur sans biais de $p$.

$$E\left(\overline{X_n}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} p = \frac{1}{n} \times np = p.$$

**Définition 10.15 –** Soient $\theta$ un réel, $X$ une variable aléatoire dont la loi dépend d'un paramètre $\theta$ et $T_n$ un estimateur de $\theta$. 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 10.16**

Soient $\theta$ un réel, $X$ une variable aléatoire dont la loi dépend d'un paramètre $\theta$ et $T_n$ un estimateur de $\theta$ qui admet une variance. On a

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

En particulier, **si $T_n$ est un estimateur sans biais de $\theta$**, 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, nous allons construire un intervalle aléatoire qui permette de "recouvrir" avec une certaine fiabilité, la vraie valeur du paramètre estimé.

**Définition 10.17 –** Soient $\theta$ un réel, $X$ une variable aléatoire dont la loi dépend d'un paramètre $\theta$ et $U_n$ et $V_n$ deux estimateurs de $\theta$. Soit $\alpha \in [0,1]$ un réel.
On dit que $[U_n, V_n]$ est un **intervalle de confiance** de $\theta$ au niveau de confiance $1-\alpha$ si

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