Chapitre 19 : Estimation

Il arrive que l'on connaisse la loi d'une variable aléatoire X, mais qu'on ignore l'une de ses caractéristiques (paramètre de la loi, espérance, variance, ...).
Dans ce chapitre, on note θ cette caractéristique inconnue et on donne des méthodes pour estimer la valeur de θ.

I)Echantillon

Déf : soit X une variable aléatoire de caractéristique inconnue θ.
Soit $n \geq 1$ un entier. On appelle n-échantillon de X tout n-uplet $(X_1, ..., X_n)$ de variables aléatoires telles que :

• $X_1, ..., X_n$ sont mutuellement indépendantes.

• $\forall k \in [\![1, n]\!]$, X_k a même loi que X.

Déf : soit $(X_1, ..., X_n)$ un n-échantillon de X.
On appelle réalisation de cet n-échantillon tout n-uplet $(x_1, ..., x_n)$ de réels tels que pour tout $k \in [\![1, n]\!]$, x_k est une valeur prise par X_k.

II)Estimateur

Déf : soit X une variable aléatoire de caractéristique inconnue θ.
Soit $(X_1, ..., X_n)$ un n-échantillon de X.
On appelle estimateur de θ, toute variable aléatoire T_n de la forme
$T_n = f(X_1, ..., X_n)$ où f est une fonction de n variables qui ne dépend pas de θ.

On appelle estimation ponctuelle de θ, le nombre réel $f(x_1, ..., x_n)$ où $(x_1, ..., x_n)$ est une réalisation du n-échantillon $(X_1, ..., X_n)$.

Exemples
Il existe une infinité d'estimateurs de θ, par exemple :

$T_n = X_1 + ... + X_n$, $T_n = X_1 ... X_n$, $T_n = \dfrac{X_1 + ... + X_n}{n}$.

La difficulté est de choisir un estimateur T_n adapté de sorte que l'estimation ponctuelle de θ qu'il fournit soit proche de la valeur exacte de θ.

III)Moyenne empirique

Déf : soit $(X_1, ..., X_n)$ un n-échantillon de X.
On appelle moyenne empirique du n-échantillon $(X_1, ..., X_n)$ la variable aléatoire, notée $\overline{X_n}$, définie par :

$$\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Théorème 1 (à savoir redémontrer)
Soit X une variable aléatoire d'espérance m et de variance σ^2.
Soit $(X_1, ..., X_n)$ un n-échantillon de X.
Soit $\overline{X_n}$, la moyenne empirique de l'échantillon.

1)$E(\overline{X_n}) = m$,

2)$V(\overline{X_n}) = \dfrac{\sigma^2}{n}$.

Exercice 1
Soit X une variable aléatoire d'espérance m et de variance σ^2.
Soit $(X_1, ..., X_n)$ un n-échantillon de X et soit $\overline{X_n}$, sa *moyenne empirique*.
Soit T_n la variable aléatoire définie par :

$$T_n = \frac{1}{n}\sum_{k=1}^{n} \left(X_k - \overline{X_n}\right)^2.$$

1)Justifier que T_n est un estimateur de σ^2.
(T_n s'appelle *variance empirique*).

2)a)Montrer que $T_n = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \left(\overline{X_n}\right)^2$.

b)Préciser $E\left(\overline{X_n}\right)$ et $V\left(\overline{X_n}\right)$ en fonction de m, n et σ.

c)Montrer que $\forall k \in [\![1, n]\!],\ E\left(X_k^2\right) = \sigma^2 + m^2$ et $E\left(\overline{X_n}^2\right) = \frac{\sigma^2}{n} + m^2$.

d)En déduire que $E(T_n) = \frac{n-1}{n}\sigma^2$. Interpréter quand n est grand.

IV)Intervalle de confiance

Déf : soit X une variable aléatoire de caractéristique θ inconnue que l'on cherche à estimer et soit $(X_1, ..., X_n)$ un n-échantillon de X.
Soient $U_n = f(X_1, ..., X_n)$ et $V_n = g(X_1, ..., X_n)$ deux estimateurs de θ.
Soit α un réel fixé de $]0, 1[$.
On dit que l'intervalle $[U_n, V_n]$ est un intervalle de confiance de θ au niveau de confiance $1 - \alpha$ (ou au risque α) si

$$P(U_n \leq \theta \leq V_n) \geq 1 - \alpha.$$

Exercice 2
Soit X une variable aléatoire d'espérance m inconnue et de variance σ^2.
Soit $(X_1, ..., X_n)$ un n-échantillon de X. Soit T_n la variable aléatoire définie par :

$$T_n = \frac{2}{n(n+1)}\sum_{k=1}^{n} kX_k.$$

1)Montrer que $E(T_n) = m$.

2)Montrer que $V(T_n) = \frac{(4n+2)\sigma^2}{3n^2+3n}$.

3)a)Al'aide de l'inégalité de Bienaymé-Tchébychev, montrer que

$$\forall \epsilon > 0,\ P(|T_n - m| < \epsilon) \geq 1 - \frac{(4n+2)\sigma^2}{(3n^2+3n)\epsilon^2}.$$

b)Conclure que $\forall \epsilon > 0,\ P(|T_n - m| \leq \epsilon) \geq 1 - \frac{(4n+2)\sigma^2}{(3n^2+3n)\epsilon^2}$.

4)A l'aide de la question 3)b), déterminer une valeur de n pour laquelle $[T_n - \sigma, T_n + \sigma]$ est un intervalle de confiance de m au niveau de confiance 0,90.

Déf : soit X une variable aléatoire de caractéristique θ inconnue que l'on cherche à estimer et soit $(X_1, ..., X_n)$ un n-échantillon de X.
Soient $U_n = f(X_1, ..., X_n)$ et $V_n = g(X_1, ..., X_n)$ deux estimateurs de θ.
Soit α un réel fixé de $]0, 1[$.

On dit que l'intervalle $[U_n, V_n]$ est un intervalle de confiance asymptotique de θ au niveau de confiance $1 - \alpha$ si

$$\lim_{n \to +\infty} P(U_n \leq \theta \leq V_n) \geq 1 - \alpha.$$

Exercice 3
Soit X une variable aléatoire d'espérance m inconnue et de variance σ^2.
Soit $(X_1, ..., X_n)$ un n-échantillon de X.
Soient $\overline{X_n}$ et S_n les variables aléatoires définies par :

$$\overline{X_n} = \frac{1}{n} \sum_{k=1}^{n} X_k \quad \text{et} \quad S_n = \sum_{k=1}^{n} X_k = n\overline{X_n}.$$

1)Déterminer la valeur de $E(S_n)$ et $V(S_n)$.

2)a)Montrer que $\forall x > 0,\ P\left(-x \leq S_n^* \leq x\right) = P\left(\overline{X_n} - \dfrac{\sigma x}{\sqrt{n}} \leq m \leq \overline{X_n} + \dfrac{\sigma x}{\sqrt{n}}\right)$.

b)A l'aide du théorème de la limite centrée, déterminer $\lim\limits_{n \to +\infty} P\left(-x \leq S_n^* \leq x\right)$ à l'aide de Φ et de x.

c)En déduire qu'un intervalle de confiance asymptotique de m au niveau de confiance 0,95 est :

$$\left[\overline{X_n} - \frac{1,96\sigma}{\sqrt{n}}, \overline{X_n} + \frac{1,96\sigma}{\sqrt{n}}\right].$$

3)On suppose que X suit la loi de Bernoulli de paramètre p.

a)Exprimer m et σ^2 en fonction de p.

b)Montrer que $\sigma \leq \dfrac{1}{2}$.

c)En déduire que $\left[\overline{X_n} - \dfrac{1}{\sqrt{n}}, \overline{X_n} + \dfrac{1}{\sqrt{n}}\right]$ est un intervalle de confiance asymptotique de p au niveau de confiance 0,95.