## 1. Rappels de statistiques : estimation, tests et intervalles de confiance

*Objectifs : Retravailler les notions d'estimation, de tests et d'intervalles de confiance. Les exercices 1.1 à 1.3 sont à faire pendant le TD, les 1.4 et 1.5 sont à chercher de votre côté.*

**Exercice 1.1** (Estimation de la variance, moyenne inconnue)**.** Soit un échantillon i.i.d. $(X_1, \dots, X_n)$ d'espérance $\mu$ et de variance $\sigma^2 > 0$ finie, toutes les deux inconnues. On s'intéresse à l'estimation de $\sigma^2$. On note

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad \widehat{m}_2 := \frac{1}{n}\sum_{i=1}^{n} X_i^2 \,.$$

On considère l'estimateur de $\sigma^2$ suivant :

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

1. Montrer que $V_n = \widehat{m}_2 - (\overline{X})^2$.
2. $V_n$ est-il un estimateur sans biais de $\sigma^2$ ? Sinon, proposer un estimateur sans biais qu'on notera $\widehat{\sigma}^2$.

Dans toute la suite de l'exercice, on suppose de plus que les $X_i$ suivent la loi normale $\mathcal{N}(\mu, \sigma^2)$. On admet que cela implique que $K_{n-1} := \sum_{i=1}^{n}(X_i - \bar{X})^2/\sigma^2$ suit une loi du khi-deux à $n-1$ degrés de libertés, loi dont la moyenne est $n-1$ et la variance est $2(n-1)$.

3. En déduire le risque quadratique de $V_n$. Cet estimateur est-il consistant ?
4. Calculer le risque quadratique de $\widehat{\sigma}^2$. Comparer avec celui de $V_n$.

Dans la suite, on cherche à estimer $\sigma^2$ avec un estimateur de la forme

$$T_{a_n} := a_n \sum_{i=1}^{n}(X_i - \bar{X})^2,$$

où $a_n$ est une constante réelle qui peut dépendre de $n$.

5. Calculer le risque de $T_{a_n}$. Quelle est une condition nécessaire et suffisante sur $a_n$ pour que $T_{a_n}$ soit consistant ?
6. A $n$ fixé, déterminer $a_n$ tel que $T_{a_n}$ soit de risque quadratique minimal.
7. A la lumière de cet exercice, déterminer si les affirmations sont vraies ou fausses, et justifier :
   (a) Un estimateur de risque minimal est forcément de variance minimale.
   (b) Un estimateur non biaisé est de risque minimal.
   (c) Un estimateur dont la variance tend vers 0 est consistant.

**Exercice 1.2** (Intervalles de confiance dans le modèle uniforme)**.** Supposons que $(X_n)_{n\geq 1}$ sont i.i.d. de loi uniforme sur $[0, \theta]$, avec $\theta > 0$. Pour tout $n \geq 1$, on définit

$$M_n := \max(X_1, \dots, X_n) \,.$$

1. Montrer que $\left(\frac{M_n}{\theta}\right)^n$ est pivotale, et donner sa loi. *On pourra calculer $\mathbb{P}((M_n/\theta)^n \leq u)$ pour tout $u \geq 0$.*

2. Soit $\alpha \in ]0,1]$. En déduire un intervalle de confiance $I_1$ de probabilité de couverture $1-\alpha$ pour $\theta$ basé sur $M_n$.

3. Trouver un équivalent du diamètre de $I_1$ lorsque $n \to \infty$, pour un $\alpha$ fixé dans $]0,1]$.

4. Montrer, en étudiant la convergence simple de la fonction de répartition, que

$$n(1 - M_n/\theta) \xrightarrow[n\to\infty]{(d)} \mathrm{Exp}(1).$$

5. Calculer *explicitement* le quantile d'ordre $\beta$ de la loi Exp(1) pour tout $\beta \in [0,1[$.

6. En déduire un intervalle de confiance asymptotique $I_2$ de probabilité de couverture $1-\alpha$ pour $\theta$.

7. Comparer le diamètre de $I_2$ à celui de $I_1$ lorsque $n \to \infty$, pour un $\alpha$ fixé dans $]0,1]$.

**Exercice 1.3** (Test gaussien, variance connue)**.** *On rappelle dans cet exercice une propriété fondamentale des v.a. gaussiennes : si $(Z_j)_{1\leq j\leq m}$ sont des v.a. réelles indépendantes et de loi respectives $(\mathcal{N}(\mu_j, \sigma_j^2))_{1\leq j\leq m}$, alors pour tous réels $\alpha_0, \alpha_1, \ldots, \alpha_m$, on a*

$$\alpha_0 + \alpha_1 Z_1 + \ldots + \alpha_m Z_m \sim \mathcal{N}\left(\alpha_0 + \sum_{j=1}^m \alpha_j \mu_j, \sum_{j=1}^m \alpha_j^2 \sigma_j^2\right).$$

Soit $(X_1, \ldots, X_{25})$ un échantillon de loi gaussienne d'espérance $\mu$ inconnue et de variance $\sigma^2 = 100$ connue.

*On donne quelques quantiles de la loi $\mathcal{N}(0,1)$:*

$$q_{0.975} \sim 1.96, q_{0.95} \sim 1.65, q_{0.9} \sim 1.28, q_{0.8} \sim 0.84,$$

*et quelques images de sa fonction de répartition $\Phi$:*

$$\Phi(1.21) \sim 0.89, \Phi(0.90) \sim 0.82, \Phi(0.53) \sim 0.70, \Phi(0.09) \sim 0.53\,.$$

1. Construire un test de niveau $\alpha = 0.10$ pour

$$\mathcal{H}_0 : \text{“}\mu = 0\text{”} \quad \text{contre} \quad \mathcal{H}_1 : \text{“}\mu = 1.5\text{”},$$

fondé sur la moyenne empirique $\bar{X} := \frac{1}{25}\sum_{i=1}^{25} X_i$, estimateur du paramètre $\mu$.

2. On observe $\bar{x} = 1$. Quelle est la décision du test ? L'erreur que l'on fait peut-être ici est-elle de première espèce ? de seconde espèce ? La calculer.

3. Déterminer la taille minimum d'un échantillon dans le même cadre que ci-dessus si l'on souhaite que le test précédent ait des erreurs de première et de seconde espèce toutes deux inférieures à 0.1.

4. Désormais on souhaite tester

$$\mathcal{H}_0 : \text{“}\mu = 2\text{”} \quad \text{contre} \quad \mathcal{H}_1 : \text{“}\mu < 2\text{”}\,.$$

Définir la région de rejet pour un niveau $\alpha$ donné. Exprimer la puissance du test à l'aide de la fonction $\Phi$, et commenter la dépendance de la puissance en fonction de $\mu$, $n$ et $\sigma$.

**Exercice 1.4** (Des questions d'identifiabilité)**.** On considère un modèle dans lequel l'observation $X$ est une différence $X = Y - Z$ avec $Y, Z$ deux variables gaussiennes indépendantes, de moyennes respectives $\mu_1, \mu_2$ et de variances respectives $\sigma_1^2, \sigma_2^2$, toutes inconnues.

1. Ce modèle est-il identifiable ?

2. Supposons dans cette question que $\mu_2 = 3\mu_1 + 1$ et $\sigma_2 = 2\sigma_1$. Cela rend-il le modèle identifiable ?

3. Qu'en est-il d'un modèle où $X = Y - Z$ avec $Y, Z$ deux variables exponentielles indépendantes à paramètres inconnus ? *On pourra chercher une interprétation géométrique aux équations pour l'espérance et la variance de $X$.*

4. Qu'en est-il d'un modèle où $X = \alpha(Y - Z)$ avec $Y, Z$ deux variables exponentielles indépendantes à paramètres inconnus, et $\alpha \in \mathbb{R}$ ? Et si $\alpha > 0$ ?

**Exercice 1.5** (Estimateur sans biais pour une Bernoulli)**.** On considère un échantillon $(X_1, \ldots, X_n)$ i.i.d. de loi de Bernoulli $\mathcal{B}(p)$.

1. Montrer que $\bar{X}$ (la moyenne emprique) est un estimateur sans biais de $p$.

2. Soit $g(\bar{X})$ un autre estimateur sans biais de $p$ qui est une fonction (mesurable) de $\bar{X}$. Montrer que pour tout $x \in \mathbb{R}^+$ :

$$\sum_{k=0}^{n} \left( g\left(\frac{k}{n}\right) - \frac{k}{n} \right) \binom{n}{k} x^k = 0$$

et en déduire que $\bar{X}$ est le seul estimateur sans biais de $p$ qui est une fonction de $\bar{X}$.