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 μ et de variance $\sigma^2 > 0$ finie, toutes les deux inconnues. On s'intéresse à l'estimation de σ^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 σ^2 suivant :

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

1. Montrer que $V_n = \widehat{m}_2 - (\overline{X})^2$. *Solution. Par simple calcul, on a*

$$V_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\bar{X}\cdot\bar{X} + \bar{X}^2 = \widehat{m}_2 - (\bar{X})^2.$$

2. V_n est-il un estimateur sans biais de σ^2 ? Sinon, proposer un estimateur sans biais qu'on notera $\widehat{\sigma}^2$. *Solution. Calculons le biais de V_n. Reprenons l'expression de l'énoncé :*

$$\begin{aligned}\mathbb{E}[V_n] &= \mathbb{E}[\widehat{m}_2] - \mathbb{E}[(\overline{X})^2] \\ &= \frac{1}{n}\sum_{i=1}^{n}(\mathrm{Var}(X_i) + \mathbb{E}[X_i]^2) - (\mathrm{Var}(\overline{X}) + \mathbb{E}[\overline{X}]^2) \\ &= \sigma^2 + \mu^2 - \frac{1}{n^2}(n\sigma^2) - \mu^2 = \frac{n-1}{n}\sigma^2,\end{aligned}$$

et le biais de V_n vaut donc $\mathbb{E}[V_n] - \sigma^2 = -\frac{1}{n}\sigma^2 < 0$, V_n est donc biaisé (avec un biais légèrement négatif). En posant :

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

on obtient un estimateur sans biais.

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 ? *Solution. On a le biais, il nous manque la variance. $V_n = \frac{\sigma^2}{n}K_{n-1}$, donc $\mathrm{Var}(V_n) = \frac{\sigma^4}{n^2}\cdot 2(n-1)$. D'où le risque quadratique :*

$$R(V_n) = \mathrm{Var}(V_n) + \big(\mathbb{E}[V_n] - \sigma^2\big)^2 = \frac{2(n-1)}{n^2}\sigma^4 + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 = \frac{2n-1}{n^2}\sigma^4.$$

On a bien $R(V_n) \to 0$ quand $n \to \infty$: V_n est consistant.

4. Calculer le risque quadratique de $\widehat{\sigma}^2$. Comparer avec celui de V_n. *Solution. On a* $\widehat{\sigma}^2 = \frac{\sigma^2}{n-1}K_{n-1}$ *donc* $\mathrm{Var}(\widehat{\sigma}^2) = \frac{2\sigma^4}{n-1}$, *et le biais étant nul (question 2), on a* $R(\widehat{\sigma}^2) = \frac{2\sigma^4}{n-1}$. *Comparons les deux risques :*

$$R(\widehat{\sigma}^2) - R(V_n) = \frac{2\sigma^4}{n-1} - \frac{2n-1}{n^2}\sigma^4 = \frac{3n-1}{(n-1)n^2}\sigma^4 > 0.$$

Donc V_n *domine* $\widehat{\sigma}^2$.

Dans la suite, on cherche à estimer σ^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 ? *Solution. Pour* $a \in \mathbb{R}$, *on a* $T_a = a\sigma^2 K_{n-1}$, *et donc* $\mathbb{E}[T_a] = a\sigma^2(n-1)$ *et* $\mathrm{Var}(T_a) = 2a^2\sigma^4(n-1)$. *Donc le risque de* T_a *vaut:*

$$R(T_a) = 2a^2\sigma^4(n-1) + \left(a(n-1)\sigma^2 - \sigma^2\right)^2.$$

On voit que ce risque tend vers 0 *ssi* $a_n \sim 1/n$ *quand* $n \to \infty$.

6. A n fixé, déterminer a_n tel que T_{a_n} soit de risque quadratique minimal. *Solution. Cela revient à minimiser* $R(T_a)/\sigma^4 = 2a^2(n-1) + (a(n-1)-1)^2$, *dont le minimum est atteint pour* $a = 1/(n+1)$. *Le risque de* $(T_{1/(n+1)})$ *vaut alors* $R(T_{1/(n+1)}) = \frac{2\sigma^4}{n+1}$, *qui est encore plus faible que celui de* V_n *ou de* $\widehat{\sigma}^2$.

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. *Solution. Faux. Par exemple,* $T_{1/(n+1)}$ *a un risque plus faible que* $T_{1/(n+2)}$, *mais ce dernier a une variance plus faible.*

   (b) Un estimateur non biaisé est de risque minimal. *Solution. Faux.* $\widehat{\sigma}^2$ *est sans biais mais a un risque plus grand que* $T_{1/(n+1)}$ *qui est biaisé.*

   (c) Un estimateur dont la variance tend vers 0 est consistant. *Solution. Faux. Par définition, il faut aussi que le biais tende vers* 0.

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

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

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* $u \geq 0$. *Solution. Trouvons la loi de* $T := \left(\frac{M_n}{\theta}\right)^n$. *Remarquons que presque sûrement,* $0 \leq T \leq 1$. *Pour tout* $0 \leq u \leq 1$, $\mathbb{P}(T \leq u) = \mathbb{P}(M_n \leq u^{1/n}\theta) = (u^{1/n}\theta/\theta)^n = u$, *donc* $T \sim \mathrm{Unif}([0,1])$, *loi qui ne dépend pas de* θ *: elle est pivotale.*

2. Soit $\alpha \in]0,1]$. En déduire un intervalle de confiance I_1 de probabilité de couverture $1-\alpha$ pour θ basé sur M_n. *Solution. On applique la méthode pivotale : pour tout* $\theta > 0$, $1-\alpha = \mathbb{P}(T \in [\alpha, 1]) = \mathbb{P}(\theta \in [M_n, M_n\alpha^{-1/n}])$, *ce qui donne l'intervalle de confiance* $I_1 = [M_n, M_n\alpha^{-1/n}]$ *avec les propriétés désirées.*

3. Trouver un équivalent du diamètre de I_1 lorsque $n \to \infty$, pour un α fixé dans $]0,1]$. *Solution. Le diamètre de I_1 est $M_n(\alpha^{-1/n} - 1) \sim \frac{M_n}{n} \log(1/\alpha)$.*

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

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

Solution. Avec – un moins – la fonction de répartition (encore une fois !) on obtient pour tout $u \geq 0$,

$$\mathbb{P}(n(1 - M_n/\theta) \geq u) = \mathbb{P}(M_n \leq \theta - u\theta/n) = \left(\frac{\theta - u\theta/n}{\theta}\right)^n = (1 - u/n)^n \to e^{-u},$$

ce qui prouve la convergence désirée.

5. Calculer *explicitement* le quantile d'ordre β de la loi Exp(1) pour tout $\beta \in [0,1[$. *Solution. Si $X \sim \text{Exp}(1)$ alors pour tout $t \geq 0$, $F_X(t) = 1 - e^{-t}$, et $\beta = F_X(t) \iff t = \log(1/(1-\beta))$.*

6. En déduire un intervalle de confiance asymptotique I_2 de probabilité de couverture $1-\alpha$ pour θ. *Solution. D'après les questions 4 et 5, pour tout $\alpha \in]0,1[$, on a, lorsque $n \to \infty$*

$$\mathbb{P}\left(M_n \leq \theta \leq \frac{M_n}{1 - \frac{1}{n}\log(1/\alpha)}\right) = \mathbb{P}\left(n(1 - M_n/\theta) \leq \log(1/\alpha)\right) \to 1 - \alpha$$

on propose donc $I_2 = \left[M_n, \frac{M_n}{1-\frac{1}{n}\log(1/\alpha)}\right]$

7. Comparer le diamètre de I_2 à celui de I_1 lorsque $n \to \infty$, pour un α fixé dans $]0,1]$. *Solution. On obtient exactement le même équivalent.*

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 μ 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 Φ:

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

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 μ. *Solution. On suit les étapes classiques de construction d'un test (cf cours). On les détaille ci-après :*

– *Modèle :* $(\mathbb{R}^{25}, \mathcal{N}(\mu, 100)^{\otimes 25})$.

– *Hypothèses :* $\mathcal{H}_0 : \mu = 0$ *contre* $\mathcal{H}_1 : \mu = 1.5$.

– *Statistique de test :*

$$T = \frac{\bar{X} - 0}{\sigma/\sqrt{n}} = \frac{\bar{X}}{2} \sim \mathcal{N}(0,1) \text{ sous } \mathcal{H}_0.$$

– *Région de rejet :* $\mathcal{R} = \{T > q_{0.9} \approx 1.28\}$, *qui a probabilité* $\alpha = 0.10$ *sous* $\mathcal{H}_0$. *La forme de la région de rejet est cohérente avec l'hypothèse* $\mathcal{H}_1$ *(on veut tester si* μ *est "plus grand").*

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. *Solution. La valeur observée est* $t^{obs} = \frac{1}{2} = 0.5 < 1.28$, *donc on ne rejette pas* $\mathcal{H}_0$. *Si on fait une erreur en ne rejettant pas* $\mathcal{H}_0$, *c'est qu'on a pas détecté* $\mathcal{H}_1$, *c'est donc une erreur de seconde espèce. Calculons-la. Sous* $\mathcal{H}_1$, *on a* $T = \frac{\bar{X}}{2} \sim \mathcal{N}(0.75, 1)$, *donc l'erreur de seconde espèce est :*

$$\beta = \mathbb{P}_1(T \leq 1.28) = \mathbb{P}_1(T - 0.75 \leq 1.28 - 0.75) = \mathbb{P}(Z \leq 0.53) \approx 0.7,$$

où Z *est une v.a. de loi* $\mathcal{N}(0,1)$. *Il y a donc environ 70% de probabilité de ne pas rejeter* $\mathcal{H}_0$ *alors que* $\mathcal{H}_1$ *est vraie. Ce test n'est pas très puissant.*

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. *Solution. Rappelons que la statistique de test s'écrit* $T = \frac{1}{10}\bar{X}\sqrt{n}$. *Sous* $\mathcal{H}_0$, $T \sim \mathcal{N}(0,1)$ *et le test est bien de niveau* 0.10 *par construction dans la question 1. Pour l'erreur de seconde espèce, on veut* $\mathbb{P}_1(T \leq q_{0.9}) \leq 0.1$, *et sous* $\mathcal{H}_1$, *après calcul,* $T \sim \mathcal{N}(0.15\sqrt{n}, 1)$. *On a* $\mathbb{P}_1(T \leq q_{0.9}) \leq 0.1$ *ssi* $q_{0.9} - 0.15\sqrt{n} \leq q_{0.1} = -1.28$. *(On utilisera la symétrie des quantile gaussiens!). Cela équivaut après calcul à* $n \geq 291.27$. *Il faut donc au minimum un échantillon de taille* $n = 292$.

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 α donné. Exprimer la puissance du test à l'aide de la fonction Φ, et commenter la dépendance de la puissance en fonction de μ, n et σ. *Solution. On prend une statistique de test proche de celle du début de l'exercice :* $T = \frac{\bar{X}-2}{\sigma/\sqrt{n}}$, *qui a loi* $\mathcal{N}(0,1)$ *sous* $\mathcal{H}_0$. *Vu la forme du test (on veut tester si* μ *est petit), on prendra logiquement comme région de rejet au niveau* α

$$\mathcal{R} = \{T < q_\alpha\}\,.$$

Calculons la puissance de ce test. Elle va dépendre bien sûr de μ. *Sous* $\mathcal{H}_1$, X_1 *est de moyenne* $\mu < 2$ *et* $T \sim \mathcal{N}\left(\frac{\mu-2}{\sigma/\sqrt{n}}, 1\right)$, *donc la puissance est donnée par*

$$\pi(\mu) = \mathbb{P}\left(\frac{\mu - 2}{\sigma/\sqrt{n}} + Z < q_\alpha^*\right) = \Phi\left(q_\alpha^* + \sqrt{n}\frac{2-\mu}{\sigma}\right)\,.$$

On remarque tout d'abord que $\pi(\mu) \to 1$ *lorsque* $n \to \infty$*: le test est asymptotiquement de puissance 1 (logique, on va avoir plein d'information pour distinguer les hypothèses). Plus généralement, la puissance augmente avec* n *(logique) et diminue avec* σ *(logique*

aussi : pourquoi?). Lorsque $\mu = 2$, on retrouve l'erreur de type 1, qui est exactement de α.

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 μ_1, μ_2 et de variances respectives σ_1^2, σ_2^2, toutes inconnues.

1. Ce modèle est-il identifiable ? *Solution. On a $Y \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $Z \sim \mathcal{N}(\mu_2, \sigma_2^2)$. On a $X \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$, et donc, en raison de l'invariance par translation, le modèle n'est pas identifiable, par exemple on peut arbitrairement modifier μ_1, μ_2 sans changer leur différence, et le statisticien ne saura pas faire la différence.*

2. Supposons dans cette question que $\mu_2 = 3\mu_1 + 1$ et $\sigma_2 = 2\sigma_1$. Cela rend-il le modèle identifiable ? *Solution. Si $Y \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $Z \sim \mathcal{N}(\mu_2, \sigma_2^2)$, alors $X \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$, il y a donc deux degrés de liberté supplémentaires dans le modèle. Sous les hypothèses en plus, $X \sim \mathcal{N}(1 - 2\mu_1, 5\sigma_1^2)$, les moments d'ordre 1 et 2 suffisent pour identifier le modèle.*

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. Solution. On obtient un système toujours soluble (il est plus simple de raisonner avec les inverses des paramètres). En posant $b = 1/\lambda$ et $c = 1/\mu$, les équations sont $b - c = \mathbb{E}[X] =: A$ et $b^2 + c^2 = \mathrm{Var}(X) =: B$. Il s'agit de trouver un point (b, c) sur le cercle centré d'équation $x^2 + y^2 = B$ à l'intersection avec la droite $y = x - A$. On voit facilement que si cette droite et ce cercle s'intersectent, ils le font au plus une fois dans le quadrant positif ; comme $b = 1/\lambda$ et $c = 1/\mu$ doivent être des solutions, il y a une solution unique. Le modèle est identifiable.*

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 $\alpha > 0$? *Solution. Quand $\alpha \in \mathbb{R}$, il existe une symétrie $(\alpha, \lambda, \mu) \mapsto (-\alpha, \mu, \lambda)$, donc le modèle n'est pas identifiable. Lorsque $\alpha > 0$, l'espérance vaut $\alpha(b - c)$ et la variance vaut $\alpha^2(b^2 + c^2)$. Les quantités αb et αc peuvent être déterminées de façon unique à partir de ces deux équations. Mais pas davantage : en effet, $\alpha Y \sim \mathrm{Exp}(\lambda/\alpha)$, il y a donc une symétrie $(\alpha, \lambda, \mu) \mapsto (1, \lambda/\alpha, \mu/\alpha)$.*

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 $\bar{X}$. *Solution. Il est clair que $\bar{X} \sim \frac{1}{n}\mathrm{Bin}(n, p)$. Par hypothèse, $g(\bar{X})$ est un estimateur sans biais de p et donc*

$$\mathbb{E}_p[g(\bar{X})] = \sum_{k=0}^{n} g\left(\frac{k}{n}\right) \binom{n}{k} p^k (1-p)^{n-k} = p$$

D'autre part, en écrivant $\mathbb{E}[\mathrm{Bin}(n,p)/n] = p$*, il vient que*

$$p = \frac{1}{n}\sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k}$$

On a donc

$$\sum_{k=0}^{n}\left(g\left(\frac{k}{n}\right)-\frac{k}{n}\right)\binom{n}{k}p^k(1-p)^{n-k} = 0$$

et ceci pour tout $p \in [0,1]$*. En posant* $x = \frac{p}{1-p}$ *(valable pour* $p \in [0,1[$*), alors pour tout* $x \in \mathbb{R}^+$*,*

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

Cette égalité est un polynôme de degré n *nul pour une infinité de valeurs de* $x \in \mathbb{R}^+$*, donc il est identiquement nul. Par linéarité des coefficients, pour tout* $1 \leq k \leq n$*,*

$$g\left(\frac{k}{n}\right) = \frac{k}{n}$$

d'où $g(\bar{X}) = \bar{X}$*. Ainsi,* $\bar{X}$ *est l'unique estimateur sans biais de* p *qui soit fonction de* $\bar{X}$*.*