Feuille de TP n°9
Régions de confiance en modélisation

1 Estimation et région de confiance

On cherche à estimer un paramètre inconnu θ d'une loi de probabilité P_θ sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$. On se place dans le cadre paramétrique avec $\theta \in \Theta$ où $\Theta \subset \mathbb{R}^d$. Une idée naturelle est d'estimer θ à partir d'un n-échantillon $(X_1, \ldots, X_n)$ de loi P_θ, à valeurs dans un espace mesurable $(E, \mathcal{E})$.

Définition 1.1 (Estimateurs). *On appelle* estimateur *de θ, basé sur les observations $(X_1, \ldots, X_n)$, tout vecteur aléatoire $\widehat{\theta}_n$ défini sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, de la forme*

$$\widehat{\theta}_n = h(X_1, \ldots, X_n)$$

où h est une application mesurable définie sur $(E, \mathcal{E})$ et à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

Remarque 1.2. $\widehat{\theta}_n$ ne dépend que des observations $(X_1, \ldots, X_n)$ et à partir de $(X_1, \ldots, X_n)$, on peut construire plusieurs estimateurs de θ. Afin de choisir le « meilleur estimateur » de θ, on dispose de différents critères de qualité.

Définition 1.3 (Propriétés des estimateurs). *Soit $\widehat{\theta}_n$ un estimateur de θ. On appelle* biais *et* risque quadratique *de $\widehat{\theta}_n$ les quantités*

$$B_n(\theta) := \mathbb{E}_\theta\Big(\widehat{\theta}_n - \theta\Big) \quad \text{et} \quad R_n(\theta) := \mathbb{E}_\theta\Big(\parallel \widehat{\theta}_n - \theta \parallel^2\Big).$$

On dit que $\widehat{\theta}_n$ est un estimateur sans biais *de θ si $B_n(\theta) = 0$ donc $\mathbb{E}_\theta(\widehat{\theta}_n) = \theta$ pour tout $\theta \in \Theta$. $\widehat{\theta}_n$ est un* estimateur consistant *de θ si $\widehat{\theta}_n \to \theta$ en probabilité, pour tout $\theta \in \Theta$ et $\widehat{\theta}_n$ est un* estimateur fortement consistant *de θ si $\widehat{\theta}_n \to \theta$ p.s. pour tout $\theta \in \Theta$.*

Définition 1.4 (Région de confiance). *Pour un paramètre inconnu $\theta \in \Theta$, on appelle* région de confiance *pour θ, de* niveau de confiance $1 - \alpha$ *avec $0 < \alpha < 1$, tout sous-ensemble mesurable aléatoire $\mathcal{C} = \mathcal{C}(X_1, \ldots, X_n)$ de $\mathbb{R}^d$, dépendant des observations $(X_1, \ldots, X_n)$, telle que $\mathbb{P}(\theta \in \mathcal{C}) \geqslant 1 - \alpha$. Si $d = 1$, on parle alors d'*intervalle de confiance.

2 Intervalle de confiance d'une moyenne

Soit $(X_1, \ldots, X_n)$ un n-échantillon de moyenne m et de variance σ^2. On cherche à construire un intervalle de confiance pour m en utilisant la moyenne empirique et la variance empirique définies par

$$\overline{X}_n := \frac{1}{n}\sum_{k=1}^{n} X_k \quad \text{et} \quad S_n^2 := \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2.$$

Ici, $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, m joue le rôle de θ, $\Theta = \mathbb{R}$, et $\mathrm{P}_\theta = \delta_\theta * \mathrm{P}$ où P est la loi commune des $X_1, \ldots, X_n$.

2.1 Variance connue

On suppose que la variance σ^2 est connue alors que la variance σ^2 est connue. La moyenne empirique $\overline{X}_n$ est un estimateur sans biais et fortement consistant de m. De plus, on a

$$\sqrt{n}\left(\frac{\overline{X}_n - m}{\sigma}\right) \xrightarrow[n\to+\infty]{\mathcal{L}} Z$$

avec $Z \sim \mathcal{N}(0,1)$. En raison de la symétrie de la loi normale, on a un intervalle de confiance symétrique

$$I = \left[\overline{X}_n - a\frac{\sigma}{\sqrt{n}}, \overline{X}_n + a\frac{\sigma}{\sqrt{n}}\right]$$

avec, pour un niveau de confiance $1-\alpha$ donné, $\mathbb{P}(|Z| \leqslant a) = 1-\alpha$. D'autre part, pour un n-échantillon de loi normale $\mathcal{N}(m,\sigma^2)$, on a

$$\sqrt{n}\left(\frac{\overline{X}_n - m}{\sigma}\right) \sim \mathcal{N}(0,1).$$

On retrouve donc le même intervalle de confiance que ci-dessus.

2.2 Variance inconnue

On suppose que la variance σ^2 est inconnue alors que la moyenne m est connue. La quantité S_n^2 est un estimateur sans biais et fortement consistant de σ^2. De plus, on a

$$\sqrt{n}\left(\frac{\overline{X}_n - m}{S_n}\right) \xrightarrow[n\to+\infty]{\mathcal{L}} Z$$

avec $Z \sim \mathcal{N}(0,1)$. On a donc un intervalle de confiance symétrique

$$I = \left[\overline{X}_n - a\frac{S_n}{\sqrt{n}}, \overline{X}_n + a\frac{S_n}{\sqrt{n}}\right]$$

avec, pour un niveau de confiance $1-\alpha$ donné, $\mathbb{P}(|Z| \leqslant a) = 1-\alpha$. D'autre part, pour un n-échantillon de loi normale $\mathcal{N}(m,\sigma^2)$, l'intervalle de confiance ci-dessus reste encore valable. Cependant, on peut affiner cet intervalle de confiance. En effet, on a $(n-1)S_n^2 \sim \sigma^2\chi^2(n-1)$ et

$$\sqrt{n}\left(\frac{\overline{X}_n - m}{S_n}\right) \sim t(n-1).$$

Comme la loi de Student est symétrique, on a un intervalle de confiance symétrique

$$I = \left[\overline{X}_n - a\frac{S_n}{\sqrt{n}}, \overline{X}_n + a\frac{S_n}{\sqrt{n}}\right]$$

avec, pour un niveau de confiance $1-\alpha$ donné, $\mathbb{P}(|T| \leqslant a) = 1-\alpha$ où $T \sim t(n-1)$.

3 Intervalle de confiance d'une proportion

On considère une population contenant deux types d'individus A et B. On cherche à construire un intervalle de confiance de la proportion p d'individus de type A. Pour ce faire, on effectue un sondage sur un échantillon de n individus. Pour $1 \leqslant k \leqslant n$, soit X_k la v.a. prenant la valeur 1 si le k^{e} individu répond qu'il est de type A et la valeur 0 sinon. La suite $(X_1,\ldots,X_n)$ est donc un n-échantillon de loi de Bernoulli $\mathcal{B}(p)$. Ici, p joue le rôle de θ, $\Theta = [0,1]$ et $\mathrm{P}_\theta = \mathcal{B}(p)$. La moyenne empirique $\overline{X}_n$ est un estimateur sans biais et fortement consistant de p. De plus, on a

$$\sqrt{n}\frac{\overline{X}_n - p}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \xrightarrow[n\to+\infty]{\mathcal{L}} Z$$

avec $Z \sim \mathcal{N}(0,1)$. On a donc un intervalle de confiance symétrique

$$I = \left[\overline{X}_n - a\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \overline{X}_n + a\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right]$$

avec, pour un niveau de confiance $1-\alpha$ donné, $\mathbb{P}(|Z| \leqslant a) = 1-\alpha$.

4 Intervalle de confiance d'une variance – cas gaussien

On suppose maintenant que $(X_1, \dots, X_n)$ est un n-échantillon de loi normale $\mathcal{N}(m, \sigma^2)$. On cherche à construire un intervalle de confiance pour σ^2. Ici, σ^2 joue le rôle de $\Theta = \mathbb{R}_+^*$, $\Theta = \mathbb{R}_+^*$ et $\mathrm{P}_\theta = \mathrm{Dil}_{\sqrt{\theta}}(\mathrm{P})$ où[1] P est la loi commune des $X_1, \dots, X_n$.

4.1 Moyenne connue

On suppose que la moyenne m est connue. On utilise alors

$$V_n = \frac{1}{\sigma^2} \sum_{k=1}^{n} (X_k - m)^2.$$

Comme $V_n \sim \chi^2(n)$ et que la loi du khi-deux n'est pas symétrique, on a un intervalle de confiance

$$I = \left[\frac{1}{b} \sum_{k=1}^{n} (X_k - m)^2, \frac{1}{a} \sum_{k=1}^{n} (X_k - m)^2 \right]$$

avec, pour un niveau de confiance $1 - \alpha$ donné, $\mathbb{P}(a \leqslant Z \leqslant b) = 1 - \alpha$ où $Z \sim \chi^2(n)$.

4.2 Moyenne inconnue

Si la moyenne m est inconnue, on a déjà vu que $(n-1)S_n^2 \sim \sigma^2 \chi^2(n-1)$. On a donc un intervalle de confiance

$$I = \left[\left(\frac{n-1}{b} \right) S_n^2, \left(\frac{n-1}{a} \right) S_n^2 \right]$$

avec, pour un niveau de confiance $1 - \alpha$ donné, $\mathbb{P}(a \leqslant Z \leqslant b) = 1 - \alpha$ où $Z \sim \chi^2(n-1)$.

5 Région de confiance pour la moyenne et la variance – cas gaussien

On suppose encore que $(X_1, \dots, X_n)$ est un n-échantillon de loi normale $\mathcal{N}(m, \sigma^2)$. On cherche à construire une région de confiance pour le couple (m, σ^2). Ce couple joue donc le rôle de θ, on a $\Theta = \mathbb{R} \times \mathbb{R}_+^*$ et $\mathrm{P}_\theta = \mathcal{N}(\theta_1, \theta_2) = \mathcal{N}(m, \sigma^2)$. Si

$$Z_n = \sqrt{n} \left(\frac{\overline{X}_n - m}{\sigma} \right) \quad \text{et} \quad T_n = \left(\frac{n-1}{\sigma^2} \right) S_n^2$$

alors, par le théorème de Cochran, $(Z_n, T_n) \sim \mathcal{N}(0,1) \otimes \chi^2(n-1)$. On a donc une région de confiance

$$\mathcal{C} = \left\{ \overline{X}_n - c\frac{\sigma}{\sqrt{n}} \leqslant m \leqslant \overline{X}_n + c\frac{\sigma}{\sqrt{n}} \quad \text{et} \quad \left(\frac{n-1}{b} \right) S_n^2 \leqslant \sigma^2 \leqslant \left(\frac{n-1}{a} \right) S_n^2 \right\}$$

avec, pour un niveau de confiance $1 - \alpha$ donné, $\alpha = \beta + \gamma - \beta\gamma$ et $\mathbb{P}(|Z| \leqslant c) = 1 - \beta$ où $Z \sim \mathcal{N}(0,1)$ et $\mathbb{P}(a \leqslant T \leqslant b) = 1 - \gamma$ où $T \sim \chi^2(n-1)$. Cette région de confiance est délimitée par une parabole et deux droites (faites donc un dessin !).

Exercice 5.1. Créer un code Matlab permettant de générer un n-échantillon de loi de Bernoulli $\mathcal{B}(p)$ où le nombre de réalisations n et le paramètre p sont affectés par l'utilisateur. Donner un intervalle de confiance à 95% pour p. Reproduire N fois la simulation précédente et déterminer le nombre de fois où l'intervalle de confiance proposé contient bien le véritable paramètre p.

[1] La loi $\mathrm{Dil}_\sigma(\mathrm{P})$ est celle de toute v.a. Z vérifiant $Z = \sigma X$ avec $X \sim \mathrm{P}$.

Exercice 5.2. Créer un code Matlab permettant de générer un n-échantillon de loi normale $\mathcal{N}(m, \sigma^2)$ où le nombre de réalisations n, la moyenne m et la variance $\sigma^2 > 0$ sont affectés par l'utilisateur. Déterminer des intervalles de confiance à 95% pour m et σ^2. Tracer dans le plan une région de confiance à 95% pour le couple (m, σ^2). Reproduire N fois la simulation précédente et déterminer le nombre de fois où la région de confiance proposée contient bien le couple (m, σ^2).

Exercice 5.3. La durée du processus d'atterrissage d'un avion est le temps, mesuré en secondes, qui s'écoule entre la prise en charge par la tour de contrôle jusqu'à l'immobilisation totale de l'appareil sur la piste. Afin de faire face au flux croissant des avions se posant à l'aéroport de Toulouse-Blagnac, une restructuration des services de la tour de contrôle, visant à diminuer la durée du processus d'atterrissage est réalisée. Auparavant, cette durée s'élevait en moyenne à 160 secondes. À la suite de la restructuration, une enquête, effectuée sur un échantillon de 1000 avions, a produit les résultats suivants:

Durée de l'atterrissage	[60,120[	[120,140[	[140,160[	[160,180[	[180,200[	[200,260[
Nombre d'avions	112	176	247	214	157	94

En supposant les données gaussiennes, déterminer un intervalle de confiance à 95% pour la moyenne de la durée du processus d'atterrissage ainsi que pour la variance associée. Peut-on affirmer, avec un niveau de confiance de 95%, que la durée du processus d'atterrissage a été diminuée par la restructuration ?

Exercice 5.4. Une tablette de chocolat sera qualifiée de qualité supérieure si elle contient une teneur en cacao supérieure à 430 grammes par kilogramme. On effectue un contrôle de qualité sur un échantillon de 10 tablettes de chocolat et on obtient les teneurs en cacao suivantes.

505,1	423,5	462,0	391,9	412,1	487,2	439,0	434,1	441,1	474,2

Dans le cadre gaussien, déterminer un intervalle de confiance à 95% pour la moyenne de la teneur en cacao. Le chocolat est-il de qualité supérieure ?

Exercice 5.5. Sur un échantillon de 1000 amateurs de café, 300 individus interrogés préfèrent le robusta à l'arabica. Donner un intervalle de confiance à 99% de la proportion d'individus préférant le robusta à l'arabica.

Exercice 5.6. On désire estimer le nombre N d'individus d'une espèce animale vivant sur une île. Pour ce faire, on capture 800 individus. Ces individus sont marqués, puis relâchés. Ensuite, on recapture ultérieurement 1000 animaux parmi lesquels on dénombre 250 animaux marqués. En déduire un intervalle de confiance à 95% pour N.

Exercice 5.7. Vous pouvez consulter plus d'une vingtaine de bases de données fournies par Stixbox de Matlab. Pour les utiliser, il suffit de taper la commande Matlab `getdata` fournie par la bibliothèque Stixbox. Déterminer des intervalles de confiance pour les moyennes et variances associées à ces bases de données, par exemple les bases 8, 18, 23, en utilisant la commande `test1n`.

