Chapitre 4

Estimation et intervalle de confiance

Nous abordons à présent la partie statistique du cours. L'objectif général de la statistique est de décrire / expliquer un phénomène aléatoire à partir d'un certain nombre d'observations de celui-ci. Le langage utilisé pour modéliser le phénomène aléatoire est naturellement celui de la théorie des probabilités. Le cas typique est celui-ci : on observe n fois un phénomène aléatoire de loi inconnue et on recueille ainsi des données $(x_1, \ldots, x_n)$. On fait alors l'hypothèse que les données x_i sont les réalisations de variables aléatoires indépendantes X_i de même loi que la loi inconnue $\mathbb{P}_X$, c'est-à-dire $x_i = X_i(\omega)$ où $\mathbb{P}_{X_i} = \mathbb{P}_X$. L'objet de la statistique (inférentielle) est précisément d'estimer la loi $\mathbb{P}_X$, ou plus modestement d'estimer certaines de ses caractéristiques (moyenne, variance etc.).

4.1 Estimation paramétrique

Le plus souvent, on fait l'hypothèse que la loi inconnue $\mathbb{P}_X$ appartient à une famille de lois connue, famille indexée par un ou plusieurs paramètres. Par exemple, la loi inconnue peut être une loi de Bernoulli $\mathcal{B}(p)$, pour un certain réel $p \in [0, 1]$, elle peut être un loi de Poisson $\mathcal{P}(\lambda)$ ou exponentielle $\mathcal{E}(\lambda)$ de paramètre $\lambda > 0$, ou encore une loi gaussienne $\mathcal{N}(\mu, \sigma^2)$ avec $\mu \in \mathbb{R}$ et $\sigma^2 > 0$. Dans ce cas, estimer la loi inconnue $\mathbb{P}_X$ revient alors à estimer (*i.e.* deviner) la valeur du/des paramètre(s).

Exemple 4.1.1. Dans l'exemple de la mutation d'un gène du chapitre précédent, on sait a priori que les variables X_i sont à valeurs dans l'ensemble $\{0, 1\}$ de sorte que X_i suit une loi de Bernoulli de paramètre p inconnu. Déterminer la loi des variables X_i revient donc à déterminer la valeur du paramètre $p \in [0, 1]$.

On introduit alors la notion d'estimateur du/des paramètre(s) inconnu(s), il s'agit d'une fonction des données $(x_1, \ldots, x_n)$ dont on espère qu'elle est un bonne approximation, en un sens à préciser, du/des paramètres inconnu(s).

Définition 4.1.2 (estimateur). On appelle *estimateur* de θ toute quantité $\widehat{\theta}_n$ qui est une fonction des données $(x_1, \ldots, x_n) = (X_1(\omega), \ldots, X_n(\omega))$.

Remarque 4.1.3. Attention, un estimateur est une fonction des seules données connues $(x_1, \ldots, x_n)$, mais il ne doit pas, par définition, dépendre du paramètre inconnu que l'on souhaite estimer.

Il faut maintenant préciser ce que l'on entend par "être une bonne approximation du paramètre inconnu θ". La notion de biais prend en compte le fait qu'en moyenne, l'estimateur $\widehat{\theta}_n$ est proche de la valeur théorique inconnue :

Définition 4.1.4 (estimateur sans biais)**.** Le *biais* est d'un estimateur $\widehat{\theta}_n$ de θ est la différence : $\theta - \mathbb{E}[\widehat{\theta}_n]$. Si $\mathbb{E}[\widehat{\theta}_n] = \theta$, on dira que l'estimateur $\widehat{\theta}_n$ est sans biais. Si $\lim_{n\to\infty} \mathbb{E}[\widehat{\theta}_n] = \theta$, on dira que l'estimateur $\widehat{\theta}_n$ est *asymptotiquement sans biais.*

Par ailleurs, on veut que lorsque la taille de l'échantillon de données $(x_1, \ldots, x_n)$ devient grande, l'estimateur $\widehat{\theta}_n$ soit arbitrairement proche de la valeur théorique θ.

Définition 4.1.5 (estimateur consistant)**.** On dit que l'estimateur $\widehat{\theta}_n$ de la quantité θ est *consistant* si lorsque n tend vers l'infini, $\widehat{\theta}_n$ converge en probabilité vers θ.

Exemple 4.1.6. On reprend l'exemple du taux de mutation du chapitre précédent. On souhaite estimer le paramètre p de la loi de Bernoulli $\mathcal{B}(p)$. La quantité ci-dessous est un estimateur de p :

$$\widehat{p}_n := \frac{S_n}{n} = \frac{X_1 + \ldots + X_n}{n}.$$

En effet, $\widehat{p}_n(\omega) = (x_1 + \ldots + x_n)/n$ est bien une fonction des seules variables $(x_1, \ldots, x_n)$. Par ailleurs, c'est un estimateur sans biais puisque :

$$\mathbb{E}[\widehat{p}_n] := \frac{\mathbb{E}[S_n]}{n} = \frac{\mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n]}{n} = \frac{p + \ldots + p}{n} = p.$$

Enfin, c'est un estimateur consistant puisque d'après la loi des grands nombres :

$$\widehat{p}_n := \frac{X_1 + \ldots + X_n}{n} \xrightarrow{\mathbb{P}} \mathbb{E}[X_1] = p.$$

On peut considérer de nombreux autres estimateurs de la quantité p, l'important est de garder à l'esprit que ce que l'on souhaite est approcher au mieux le paramètre p. Par exemple, $\widetilde{p}_n = X_1$ est bien une fonction des seules données. C'est donc un estimateur de p, et on peut ajouter qu'il est sans biais puisque si $X_1 \sim \mathcal{B}(p)$, on a $\mathbb{E}[X_1] = p$ et donc $\mathbb{E}[\widetilde{p}_n] = p$. En revanche, $\widetilde{p}_n$ n'est pas consistant puisqu'il ne dépend pas du nombre n de données. Au contraire, l'estimateur

$$\dot{p}_n := \frac{X_1 + \ldots + X_n}{n+1}$$

possède un biais de $p - \mathbb{E}[\dot{p}_n] = p/(n+1)$. Il est donc asymptotiquement sans biais, et d'après la loi des grands nombres, il est consistant.

4.1.1 Estimateurs empiriques

Nous introduisons maintenant une classe importante et naturelle d'estimateurs : les estimateurs empiriques. Ce sont les estimateurs construits à partir de somme de variables aléatoires et dont le comportement asymptotique peut être facilement décrit grâce à la loi des grands nombres et au théorème limite central.

Définition 4.1.7 (estimateurs empiriques)**.** On appelle moyenne empirique de l'échantillon $(x_1, \ldots, x_n) = (X_1(\omega), \ldots, X_n(\omega))$ la moyenne arithmétique

$$\widehat{m}_n = \frac{x_1 + \ldots + x_n}{n} = \frac{X_1(\omega) + \ldots + X_n(\omega)}{n}.$$

On appelle variance empirique de l'échantillon $(x_1, \ldots, x_n)$ la quantité :

$$\widehat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{m}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \widehat{m}_n^2.$$

Si les variables X_i de loi inconnue admettent des moments d'ordre un et deux, alors la loi des grands nombres assure que la moyenne et la variance empirique sont des estimateurs consistants de la moyenne m et de la variance σ^2 théoriques. En effet, d'après la loi des grands nombres, on a

$$\widehat{m}_n = \frac{X_1 + \ldots + X_n}{n} \xrightarrow{\mathbb{P}} \mathbb{E}[X_1] = m,$$

et

$$\frac{X_1^2 + \ldots + X_n^2}{n} \xrightarrow{\mathbb{P}} \mathbb{E}[X_1^2],$$

d'où

$$\widehat{\sigma}_n^2 \xrightarrow{\mathbb{P}} \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \sigma^2.$$

Exemple 4.1.8. On reprend l'exemple du nombre d'accidents envisagé au chapitre précédent. On fait l'hypothèse que les données sont des réalisations indépendantes de variables X_i de loi de Poisson $\mathcal{P}(\lambda)$ où λ est à déterminer. On a vu en cours et en TD que l'espérance et la variance d'une loi de Poisson sont $m = \mathbb{E}[X_i] = \lambda$ et $\sigma^2 = \mathrm{var}(X_i) = \lambda$. Pour estimer le paramètre λ, on peut donc naturellement choisir les estimateurs empiriques $\widehat{m}_n$ et $\widehat{\sigma}_n^2$.

Si le paramètre θ à déterminer s'écrit comme une fonction de la moyenne des variables de l'échantillon, c'est-à-dire si $\theta = g(\mathbb{E}[X])$ pour une certaine fonction continue g, alors un estimateur naturel de θ est donnée par :

$$\widehat{\theta}_n = g(\widehat{m}_n).$$

En effet, d'après la loi des grands nombres, lorsque n tend vers l'infini, on a alors

$$\widehat{\theta}_n = g(\widehat{m}_n) \xrightarrow{\mathbb{P}} g(\mathbb{E}[X]) = g(\theta),$$

autrement dit, l'estimateur $\widehat{\theta}_n$ est consistant.

Exemple 4.1.9. On recueille des données $(x_1, \ldots, x_n)$ dont on fait l'hypothèse qu'elles sont des réalisations indépendantes de variables X_i de loi uniforme sur un intervalle $[0, \theta]$ où θ est à déterminer. On a vu en cours et en TD que l'espérance d'une telle loi est $m = \mathbb{E}[X_i] = \theta/2$, autrement dit $\theta = 2\mathbb{E}[X]$. Alors un estimateur naturel de θ est donné par

$$\widehat{\theta}_n = \frac{2(x_1 + \ldots + x_n)}{n} = 2\widehat{m}_n.$$

En effet, d'après la loi des grands nombres, lorsque n tend vers l'infini, on a alors

$$\widehat{\theta}_n \xrightarrow{\mathbb{P}} 2\mathbb{E}[X] = \theta.$$

Exemple 4.1.10. On recueille des données $(x_1, \ldots, x_n)$ dont on fait l'hypothèse qu'elles sont des réalisations indépendantes de variables X_i de loi de exponentielle $\mathcal{E}(\lambda)$ où λ est à déterminer. On a vu en cours et en TD que l'espérance d'une loi exponentielle est $m = \mathbb{E}[X_i] = 1/\lambda$, autrement dit $\lambda = 1/\mathbb{E}[X]$. Alors un estimateur naturel de λ est donné par

$$\widehat{\lambda}_n = \frac{n}{x_1 + \ldots + x_n} = \frac{1}{\widehat{m}_n}.$$

En effet, d'après la loi des grands nombres, lorsque n tend vers l'infini, on a alors

$$\widehat{\lambda}_n \xrightarrow{\mathbb{P}} \frac{1}{\mathbb{E}[X]} = \lambda.$$

4.1.2 Maximum de vraissemblance

Nous venons de voir que la loi des grands nombres permet souvent de mettre en évidence des estimateurs naturels. Une autre façon de trouver de tels estimateurs est d'utiliser la méthode du maximum de vraissemblance décrite ci-dessous.

Considérons ainsi des réalisations x_i de variables aléatoires X_i indépendantes et de même loi, admettant une densité commune f_θ qui dépend du paramètre à estimer θ. Par exemple, les variables en question peuvent être des variables exponentielles de paramètre θ, de sorte que $f_\theta(x) = \theta \exp(-\theta x)$ pour $x \geqslant 0$.

Définition 4.1.11 (estimateur du maximum de vraissemblance). Étant données des variables aléatoires de loi à densité f_θ, on appelle estimateur du maximum de vraissemblance la quantité

$$\arg\max_\theta \prod_{i=1}^{n} f_\theta(x_i),$$

ou de manière équivalente

$$\arg\max_\theta \sum_{i=1}^{n} \log\left(f_\theta(x_i)\right).$$

Dans le cas des variables exponentielles, on a

$$\prod_{i=1}^{n} f_\theta(x_i) = \theta^n \exp\left(-\theta \sum_{i=1}^{n} x_i\right) = \theta^n \exp\left(-n\theta \widehat{m}_n\right)$$

où $\widehat{m}_n$ est la moyenne empirique. En prenant le logarithme, on obtient :

$$\sum_{i=1}^{n} \log\left(f_\theta(x_i)\right) = n\log(\theta) - n\theta\widehat{m}_n.$$

On cherche à trouver le maximum de cette fonction. Pour cela, on regarde quand sa dérivée par rapport à θ s'annulle. Le calcul donne :

$$\frac{\partial}{\partial\theta} n\log(\theta) - n\theta\widehat{m}_n = \frac{n}{\theta} - n\widehat{m}_n.$$

Cette expression s'annule si et seulement si $\theta = 1/\widehat{m}_n$, autrement dit :

$$\arg\max_\theta \sum_{i=1}^{n} \log\left(f_\theta(x_i)\right) = 1/\widehat{m}_n.$$

L'estimateur du maximum de vraissemblance n'est autre que $1/\widehat{m}_n$, *i.e.* on retrouve l'estimateur de l'exemple 4.1.10.

Exemple 4.1.12. On reprend l'exemple des variables uniforme sur l'intervalle $[0,\theta]$ où θ est à déterminer. La densité d'une telle variable est la fonction $f_\theta(x) = 1/\theta$ si $x \in [0,\theta]$ et zéro ailleurs. Dès lors,

$$\begin{aligned} V(\theta) := \textstyle\prod_{i=1}^{n} f_\theta(x_i) \quad &= \theta^{-n} \text{ si pour tout } i \;\; 0 \leqslant x_i \leqslant \theta, \text{ et zéro ailleurs} \\ &= \theta^{-n} \text{ si } \; 0 \leqslant \max x_i \leqslant \theta, \text{ et zéro ailleurs.} \end{aligned}$$

Le maximum de la fonction V est atteint en $\theta = \max x_i$, autrement dit, l'estimateur du maximum de vraissemblance du paramètre θ est ici donné par $\widehat{\theta}_n = \max_{i=1\ldots n} x_i$. Si l'on fixe un $\varepsilon > 0$, on a alors

$$\mathbb{P}(|\theta - \widehat{\theta}_n| > \varepsilon) = \mathbb{P}(\max X_i < \theta - \varepsilon) = \left(\frac{\theta-\varepsilon}{\theta}\right)^n \xrightarrow{n\to+\infty} 0.$$

Autrement dit, $\widehat{\theta}_n$ est un estimateur consistant de θ.

Remarque 4.1.13. Dans certains cas simples comme celui de l'estimation du paramètre d'une loi exponentielle envisagé ci-dessus, l'estimateur obtenu via la méthode du maximum de vraisemblance coïncide avec l'estimateur empirique. Ce n'est pas le cas en général comme en atteste le dernier exemple concernant la loi uniforme. Dans les cas où la maximisation de la vraisemblance est explicitement possible, et lorsqu'il diffère de l'estimateur empirique, on préfèrera l'estimateur du maximum de vraisemblance dont on peut montrer qu'il possède en général de meilleures propriétés asymptotiques.

4.2 Intervalles de confiance

Dans la section précédente, nous avons vu différentes méthodes pour estimer les paramètres d'une loi de probabilité inconnue. Nous avons par ailleurs introduit les notions de biais et de consistance qui permettent d'évaluer qualitativement un estimateur. Il arrive souvent dans la pratique que l'on veuille de plus évaluer quantitativement la qualité d'un estimateur : on peut par exemple chercher à savoir à quel vitesse (en fonction de la taille de l'échantillon) il converge vers la quantité à estimer, ou encore quelle est la probabilité de se tromper en disant que l'estimateur est proche de sa cible etc. La notion d'*intervalle de confiance*, on parle aussi de *zone de confiance*, permet précisément de quantifier la qualité d'un estimateur.

Définition 4.2.1. Soit $\alpha \in]0,1[$. On dit qu'un intervalle $I = I(X_1, \ldots, X_n)$ qui s'exprime en fonction de l'echantillon est un *intervalle de confiance* pour θ de niveau $1-\alpha$ si

$$\mathbb{P}(\theta \in I(X_1, \ldots, X_n)) = 1 - \alpha.$$

Lorsque $\mathbb{P}(\theta \in I(X_1, \ldots, X_n)) \geqslant 1-\alpha$, on parle d'intervalle de confiance de niveau $1-\alpha$ par excès.

Remarque 4.2.2. Les niveaux usuels sont 90%, 95% et 99% et correspondent respectivement à $\alpha = 10\%$, $\alpha = 5\%$ et $\alpha = 1\%$. Pour obtenir le maximum d'information, il faut s'efforcer de construire l'intervalle de confiance le moins large possible qui satisfait la condition de minoration donnée dans la définition.

Exemple 4.2.3. Considérons un $n-$échantillon $(X_1, \ldots, X_n)$ de variables aléatoires gaussiennes $\mathcal{N}(\mu, 1)$ où la moyenne μ est inconnue. Si $\widehat{X}_n$ désigne la moyenne empirique de l'échantillon, il facile de voir que la variable $Z = \sqrt{n} \times (\widehat{X}_n - \mu)$ a la même loi qu'une gaussienne $\mathcal{N}(0,1)$. Soit alors $\alpha = 5\%$, et $\beta = 1,960$ de sorte que $\mathbb{P}(|\mathcal{N}(0,1)| > \beta) = \alpha$. Alors, on a

$$\mathbb{P}(|Z| \leqslant \beta) = 1 - \alpha,$$

c'est-à-dire

$$\mathbb{P}\left(\mu \in \left[\widehat{X}_n - \frac{\beta}{\sqrt{n}}, \widehat{X}_n + \frac{\beta}{\sqrt{n}}\right]\right) = 1 - \alpha,$$

autrement dit, $I = [\widehat{X}_n - \beta/\sqrt{n}, \widehat{X}_n + \beta/\sqrt{n}]$ est un intervalle de confiance de niveau α pour le paramètre θ.

Définition 4.2.4. Soit $\alpha \in]0,1[$. On appelle intervalle de confiance asymptotique pour θ de niveau $1-\alpha$ une suite $I_n = I(X_1, \ldots, X_n)$ d'intervalles de confiance tels que

$$\lim_{n \to +\infty} \mathbb{P}(\theta \in I_n(X_1, \ldots, X_n)) = 1 - \alpha.$$

Exemple 4.2.5. On reprend les exemples du chapitre précédent sur les théorèmes limites fondamentaux. Dans le cas du taux de mutation d'un gène, d'après la loi des grands nombres, la moyenne empirique $\widehat{p}_n = S_n/n$ est un estimateur consistant du paramètre inconnu p. Soit $x_0 = 1.96$ de sorte que $\mathbb{P}(|\mathcal{N}(0,1)| > x_0) = 5\%$. D'après le théorème limite central, lorsque n tend vers l'infini, un intervalle de confiance asymptotique pour p de niveau 95% est donné par :

$$I_n := \left[\widehat{p}_n - \frac{x_0}{2\sqrt{n}},\ \widehat{p}_n + \frac{x_0}{2\sqrt{n}}\right].$$

De la même façon, dans le cas de l'estimation de la moyenne λ d'une loi de Poisson (nombre de sinistres), on a vu que si $x_0 = 2.5758$ l'intervalle suivant est un intervalle de confiance asymptotique pour λ de niveau 99% :

$$I_n = \left[\frac{S_n}{n} - \sqrt{\frac{S_n}{n^2}} \times x_0,\ \frac{S_n}{n} + \sqrt{\frac{S_n}{n^2}} \times x_0\right].$$

Exemple 4.2.6. Un sondage auprès d'un échantillon de n personnes sur leur intention de vote au second tour d'une élection indique que 46% des sondés veulent voter pour A et 54% pour B. On veut donner un intervalle de confiance asymptotique de niveau 95% de la proportion p des français qui souhaitent voter pour A. On peut modéliser les réponses des sondés (pris au hasard dans la population) par des variables aléatoires X_i de loi de Bernoulli $\mathcal{B}(p)$: $X_i = 1$ si la $i-$ème personne interrogée vote pour A, $X_i = 0$ si la $i-$ème personne interrogée vote pour B. D'après l'énoncé, la proportion de personne ayant l'intention de voter pour A, c'est-à-dire la moyenne empirique $\widehat{X}_n$ vaut 46%. Comme dans le cas du taux de mutation, si $x_0 = 1.96$ de sorte que $\mathbb{P}(|\mathcal{N}(0,1)| > x_0) = 5\%$, on montre qu'un intervalle de confiance asymptotique pour la proportion p est donné par :

$$I_n := \left[\widehat{X}_n - \frac{x_0}{2\sqrt{n}},\ \widehat{X}_n + \frac{x_0}{2\sqrt{n}}\right].$$

Si $n = 100$ on obtient ainsi l'intervalle

$$I_n = \left[0.46 - \frac{1.96}{2 \times 10}, 0.46 + \frac{1.96}{2 \times 10}\right] \approx [0.36, 0.55],$$

et l'issue de l'élection est très incertaine. Lorsque $n = 1000$, on trouve

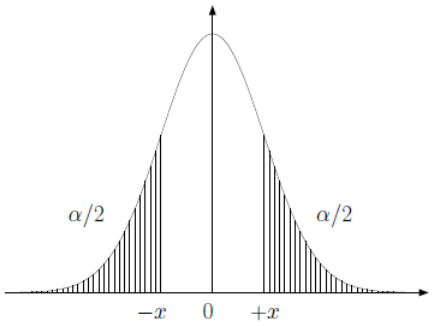
$$I_n = \left[0.46 - \frac{1.96}{2 \times \sqrt{1000}}, 0.46 + \frac{1.96}{2 \times \sqrt{1000}}\right] \approx [0.43, 0.49],$$

et avec 95 chances sur 100, on peut affirmer que le candidat A perdra l'élection.

Tables de la loi normale centrée réduite

$$2\int_x^\infty \mathrm{e}^{-y^2/2}\ \frac{dy}{\sqrt{2\pi}} = \mathbb{P}(|X| \geq x) = \alpha.$$

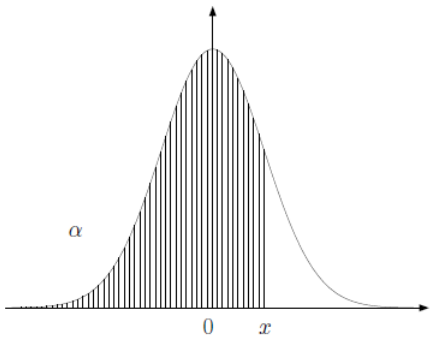
La table donne les valeurs de x en fonction de α. Par exemple $\mathbb{P}(|X| \geq 0.6280) \simeq 0.53$.



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	∞	2.5758	2.3263	2.1701	2.0537	1.9600	1.8808	1.8119	1.7507	1.6954
0.1	1.6449	1.5982	1.5548	1.5141	1.4758	1.4395	1.4051	1.3722	1.3408	1.3106
0.2	1.2816	1.2536	1.2265	1.2004	1.1750	1.1503	1.1264	1.1031	1.0803	1.0581
0.3	1.0364	1.0152	0.9945	0.9741	0.9542	0.9346	0.9154	0.8965	0.8779	0.8596
0.4	0.8416	0.8239	0.8064	0.7892	0.7722	0.7554	0.7388	0.7225	0.7063	0.6903
0.5	0.6745	0.6588	0.6433	0.6280	0.6128	0.5978	0.5828	0.5681	0.5534	0.5388
0.6	0.5244	0.5101	0.4959	0.4817	0.4677	0.4538	0.4399	0.4261	0.4125	0.3989
0.7	0.3853	0.3719	0.3585	0.3451	0.3319	0.3186	0.3055	0.2924	0.2793	0.2663
0.8	0.2533	0.2404	0.2275	0.2147	0.2019	0.1891	0.1764	0.1637	0.1510	0.1383
0.9	0.1257	0.1130	0.1004	0.0878	0.0753	0.0627	0.0502	0.0376	0.0251	0.0125

$$\int_{-\infty}^x \mathrm{e}^{-y^2/2}\ \frac{dy}{\sqrt{2\pi}} = \mathbb{P}(X \leq x) = \alpha.$$

La table suivante donne les valeurs de $1-\alpha$ pour les grandes valeurs de x.



x	2	3	4	5	6	7	8	9	10
$1-\alpha$	2.28e-02	1.35e-03	3.17e-05	2.87e-07	9.87e-10	1.28e-12	6.22e-16	1.13e-19	7.62e-24

FIGURE 4.1 – Quantiles de la loi normale centrée réduite.