Chapitre 17 :

STATISTIQUES INFÉRENTIELLES : ESTIMATION

1. Position du problème

1.1 Exemple et problématique

Exemple 1.1 À l'approche du second tour d'une élection présidentielle opposant deux candidats A et B, on interroge n individus dans la population française sur leur intention de vote. On note $X_i = 1$ lorsque le i-ième individu se prononce pour A et $X_i = 0$ sinon. On suppose que les sondés sont choisis au hasard (en excluant les abstentionnistes), de telle sorte que l'on peut modéliser les variables $X_1, \dots, X_n$ comme étant indépendantes [1] et identiquement distribuées, de loi commune une loi de Bernoulli de paramètre $\theta \in [0, 1]$ inconnu. Ce paramètre θ est la proportion de Français qui voteraient pour A si l'élection se déroulait le jour où le sondage est réalisé. Au vu des réalisations des variables aléatoires X_i, on cherche à *estimer* θ, mais aussi à savoir si le candidat A va être élu, c'est-à-dire *tester* si θ est supérieur à $\frac{1}{2}$ ou non.

L'objet de la statistique inférentielle est de répondre aux problèmes soulevés dans l'exemple précédent. Conformément au programme, on se concentrera uniquement sur le problème de l'estimation.
Il faut noter que, comme en probabilités, le hasard intervient fortement. Mais dans la théorie des probabilités, on suppose la loi connue précisément et on étudie alors le comportement d'une variable aléatoire qui suit cette loi. La démarche en statistiques est inverse : à partir de la connaissance de la variable, que peut-on dire de la loi de cette variable ?

1.2 Modèle statistique

On se donne pour objectif d'estimer certaines valeurs caractéristiques de la loi de probabilité régissant l'observation d'un nombre réel associé à une expérience aléatoire reproductible dans des conditions identiques et indépendantes.

On se place dans la situation où une connaissance partielle préalable de l'expérience aléatoire permet de modéliser la quantité étudiée comme la réalisation d'une variable aléatoire réelle X dont la loi appartient à une famille $(\nu_\theta)_{\theta\in\Theta}$ de lois de probabilités paramétrée par un ensemble $\Theta \subset \mathbb{R}^k$ ($k \geqslant 1$).

Remarques 1.2
- Par *réalisation* d'une variable aléatoire X définie sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$, on entend la valeur $x = X(\omega)$ que prend la variable X en un certain ω de l'ensemble Ω sur lequel elle est définie.
- Il existe une valeur inconnue du paramètre $\theta \in \Theta$, parfois notée θ_0 et appelée « vrai paramètre », telle que X suive la loi ν_{θ_0}.

Bien entendu, une réalisation d'une seule variable aléatoire de loi ν ne permettra pas d'obtenir beaucoup d'informations sur ν ! On est donc conduit à dégager la notion d'échantillon :

Définition 1.3 *Soient $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé, ν une loi de probabilité et $n \in \mathbb{N}^*$.*
*(i) On appelle **n-échantillon** de la loi ν toute famille $(X_1, \dots, X_n)$ de variables aléatoires X_i sur $(\Omega, \mathcal{A}, \mathbb{P})$ suivant toutes la loi ν.*
*(ii) Un n-échantillon est dit **indépendant identiquement distribué** (en abrégé, i.i.d.) lorsque les variables X_i sont de plus mutuellement indépendantes (pour la probabilité $\mathbb{P}$).*

On admet le théorème suivant, qui donne l'existence du *modèle statistique* sur lequel on travaillera dans le reste du chapitre.

Théorème 1.4 *(i) Il existe sur l'ensemble Ω des suites réelles $(x_n)_{n\in\mathbb{N}^*}$ une tribu $\mathcal{A}$ telle que pour tout $i \in \mathbb{N}^*$, l'application $X_i : (x_n)_{n\in\mathbb{N}^*} \in \Omega \longmapsto x_i$ soit une variable aléatoire.*
(ii) Pour tout $\theta \in \Theta$, il existe une mesure de probabilité $\mathbb{P}_\theta$ sur $(\Omega, \mathcal{A})$ telle que, pour tout $n \in \mathbb{N}^*$, $(X_1, \dots, X_n)$ soit un n-échantillon $\mathbb{P}_\theta$-indépendant identiquement distribué selon la loi ν_θ.

1. En toute rigueur, le nombre de sondés en faveur de A suit une loi hypergéométrique et les X_i ne sont pas indépendantes. Mais n étant négligeable devant la population française, on peut approcher cette loi par une loi binomiale et supposer les X_i indépendantes.

Remarques 1.5 • Un modèle statistique est donc un espace probabilisable muni, non pas d'une mesure de probabilité, mais d'une famille de mesures de probabilité.

• Ne connaissant pas la vraie valeur du paramètre θ, on sera amené à effectuer des calculs sous toutes les probabilités $\mathbb{P}_\theta$. On notera $\mathbb{E}_\theta$ et $\mathbb{V}_\theta$ les opérateurs espérance et variance pour la probabilité $\mathbb{P}_\theta$.

2. Estimation ponctuelle

2.1 Notion d'estimateur

Étant donnée une fonction $g : \Theta \longrightarrow \mathbb{R}$, on se donne pour objectif d'estimer $g(\theta)$, qui représentera une valeur caractéristique de la loi ν_θ telle que son espérance, sa variance, son étendue...

DÉFINITION 2.1 *(i) On appelle* ***statistique*** *toute variable aléatoire de la forme $\varphi_n(X_1, \dots, X_n)$, où φ_n est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$.*
(ii) Un ***estimateur*** *de $g(\theta)$ est une statistique $T_n = \varphi_n(X_1, \dots, X_n)$.*

Remarques 2.2 • Concrètement, c'est la réalisation $t_n = \varphi_n(x_1, \dots, x_n)$ de $T_n = \varphi_n(X_1, \dots, X_n)$ qui fournit une estimation de $g(\theta)$; on l'appelle une *estimée* de $g(\theta)$.

• Le paramètre θ étant inconnu, il est bien entendu indispensable que φ_n n'en dépende pas.

• On rencontre la notation $\widehat{g}_n$ pour désigner un estimateur de $g(\theta)$.

Exemples 2.3 (i) Si $(X_1, \dots, X_n)$ est un n-échantillon i.i.d. de loi $\mathcal{B}(p)$ de paramètre inconnu p, alors

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad T_n = \frac{2}{n(n+1)}\sum_{i=1}^{n} iX_i \quad \text{et} \quad U_n = 0$$

sont des estimateurs de p.

(ii) Si $(X_1, \dots, X_n)$ est un n-échantillon i.i.d. de loi $\mathcal{U}([a, b])$ de paramètre inconnu $(a, b) \in \mathbb{R}^2$, $a < b$, alors

$$V_n = \max_{1 \leqslant i \leqslant n} X_i - \min_{1 \leqslant i \leqslant n} X_i$$

est un estimateur de $b - a$.

Remarque 2.4 La définition précédente est très générale et ne présume en rien de la pertinence d'un estimateur. Les sections suivantes présentent différentes mesures de la qualité d'un estimateur.

2.2 Biais d'un estimateur

Soient $n \in \mathbb{N}^*$ et $T_n = \varphi_n(X_1, \dots, X_n)$ un estimateur de $g(\theta)$.

DÉFINITION 2.5 *On suppose que l'estimateur T_n admet une espérance pour chaque probabilité $\mathbb{P}_\theta$, $\theta \in \Theta$.*
(i) On appelle ***biais*** *de T_n la fonction*

$$b(T_n) : \theta \in \Theta \longmapsto b_\theta(T_n) = \mathbb{E}_\theta(T_n) - g(\theta).$$

(ii) On dit que l'estimateur T_n est ***sans biais*** *si son biais est identiquement nul c'est-à-dire :*

$$\forall \theta \in \Theta, \quad \mathbb{E}_\theta(T_n) = g(\theta).$$

Dans le cas contraire, on dit que T_n est un estimateur ***biaisé****.*

Exemple 2.6 Dans l'exemple 2.3, S_n et T_n sont des estimateurs sans biais alors que U_n et V_n sont biaisés.

Remarque 2.7 Le caractère non biaisé n'est pas indispensable pour avoir un « bon estimateur ». En effet, un estimateur faiblement biaisé mais de faible variance sera sans doute meilleur qu'un estimateur sans biais mais de forte variance, qui prendra souvent des valeurs éloignées de $g(\theta)$. La section suivante introduit un nouveau critère de qualité prenant ces considérations en compte.

2.3 Risque quadratique d'un estimateur

Soient $n \in \mathbb{N}^*$ et $T_n = \varphi_n(X_1, \dots, X_n)$ un estimateur de $g(\theta)$.

DÉFINITION 2.8 *On suppose que l'estimateur T_n admet un moment d'ordre 2 sous $\mathbb{P}_\theta$ pour tout $\theta \in \Theta$.*
*On appelle **risque quadratique** de T_n la fonction*

$$r(T_n) : \theta \in \Theta \longmapsto r_\theta(T_n) = \mathbb{E}_\theta\big[\big(T_n - g(\theta)\big)^2\big].$$

Exemple 2.9 Avec les notations de l'exemple 2.3, on a pour tout $\theta \in [0,1]$:

$$\forall n \in \mathbb{N}^*, \quad r_\theta(S_n) = \frac{1}{n}p(1-p)$$

et :

$$r_\theta(T_n) \sim \frac{4}{3n}p(1-p), \quad n \to \infty.$$

PROPOSITION 2.10 *On suppose que l'estimateur T_n admet un moment d'ordre 2 sous $\mathbb{P}_\theta$ pour tout $\theta \in \Theta$.*
(i) Si T_n est un estimateur sans biais, alors :

$$\forall \theta \in \Theta, \quad r_\theta(T_n) = \mathbb{V}_\theta(T_n).$$

(ii) Plus généralement,

$$\forall \theta \in \Theta, \quad r_\theta(T_n) = \mathbb{V}_\theta(T_n) + b_\theta(T_n)^2.$$

Remarque 2.11 Un estimateur est d'autant meilleur que son risque est faible.

2.4 Convergence d'une suite d'estimateurs

Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite d'estimateurs de $g(\theta)$ où, pour tout $n \in \mathbb{N}^*$, $T_n = \varphi_n(X_1, \dots, X_n)$ pour une fonction $\varphi_n : \mathbb{R}^n \longrightarrow \mathbb{R}$.

DÉFINITION 2.12 *On dit que $(T_n)_{n\in\mathbb{N}^*}$ est une suite d'estimateurs **asymptotiquement sans biais** de $g(\theta)$ si*

$$\forall \theta \in \Theta, \quad \lim_{n\to\infty} \mathbb{E}_\theta(T_n) = g(\theta).$$

Remarque 2.13 Par abus de langage, on dit plus simplement que T_n est un estimateur asymptotiquement sans biais de $g(\theta)$.

Exemple 2.14 Si $(X_1, \dots, X_n)$ est un n-échantillon i.i.d. de loi $\mathcal{U}([a,b])$ de paramètre inconnu $(a,b) \in \mathbb{R}^2$, $a < b$, alors $T_n = \max(X_1, \dots, X_n)$ est un estimateur asymptotiquement sans biais de b.

DÉFINITION 2.15 *On dit que T_n est **convergent** (ou **consistant**) si, pour tout $\theta \in \Theta$, la suite $(T_n)_{n\in\mathbb{N}^*}$ converge en $\mathbb{P}_\theta$-probabilité vers $g(\theta)$:*

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} \mathbb{P}_\theta\big(|T_n - g(\theta)| \geqslant \varepsilon\big) = 0.$$

Exemple 2.16 Dans l'exemple 2.14, T_n est un estimateur convergent de b.

Remarque 2.17 L'inégalité de Markov ou, dans les cas les plus simples, celle de Bienaymé-Tchebychev sont des outils intéressants pour établir qu'un estimateur est convergent.

THÉORÈME 2.18 *On suppose que pour tout $\theta \in \Theta$, la loi ν_θ admet un moment d'ordre 2 et que son espérance m_θ est de la forme $m_\theta = g(\theta)$ pour une fonction $g : \Theta \longrightarrow \mathbb{R}$.*
Dans ces conditions, la moyenne empirique

$$T_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

est un estimateur sans biais et convergent de m_θ.

PROPOSITION 2.19 *Soit $f : \mathbb{R} \longrightarrow \mathbb{R}$ une fonction continue.*
Si T_n est un estimateur convergent de $g(\theta)$, alors $f(T_n)$ est un estimateur convergent de $f\big(g(\theta)\big)$.

Remarque 2.20 Le résultat ci-dessus est encore valable lorsque f est seulement continue en $g(\theta)$ pour tout $\theta \in \Theta$.

THÉORÈME 2.21 *Si*

$$\forall \theta \in \Theta, \quad \lim_{n\to\infty} r_\theta(T_n) = 0,$$

alors T_n est un estimateur convergent de $g(\theta)$.

3. Estimation par intervalle de confiance

3.1 Notion d'intervalle de confiance

Soit $g : \Theta \longrightarrow \mathbb{R}$ une fonction. S'il existe des critères pour juger des qualités d'un estimateur ponctuel T_n de $g(\theta)$ (biais, risque, convergence), aucune certitude ne peut jamais être apportée quant au fait que l'estimation donne la vraie valeur à estimer. On peut simplement dire que si T_n est un « bon » estimateur de $g(\theta)$, il y a de fortes chances que l'estimée $t_n = T_n(\omega)$ constitue une bonne approximation de $g(\theta)$. La démarche de l'estimation par intervalle de confiance permet de quantifier cela en proposant un intervalle aléatoire qui contienne $g(\theta)$ avec une probabilité minimale donnée.

DÉFINITION 3.1 *Soient $U_n = \varphi_n(X_1, \dots, X_n)$ et $V_n = \psi_n(X_1, \dots, X_n)$ deux statistiques telles que $U_n \leqslant V_n$ $\mathbb{P}_\theta$-presque sûrement pour tout $\theta \in \Theta$.*
*Pour $\alpha \in [0,1]$ donné, on dit que $[U_n, V_n]$ est un **intervalle de confiance** de $g(\theta)$ au **niveau de confiance** $1-\alpha$ si :*

$$\forall \theta \in \Theta, \quad \mathbb{P}_\theta\big([U_n \leqslant g(\theta) \leqslant V_n]\big) \geqslant 1-\alpha.$$

*Le réel α est appelé **risque** de l'intervalle de confiance.*

Remarques 3.2
- Si u_n et v_n sont les réalisations des statistiques U_n et V_n, alors on peut affirmer que $u_n \leqslant g(\theta) \leqslant v_n$ avec le risque α de commettre une erreur.
- Bien entendu, on recherchera un intervalle de confiance le plus petit possible.

Exemple 3.3 (Estimation par intervalle du paramètre d'une loi de Bernoulli grâce à l'inégalité de Bienaymé-Tchebychev) Si $(X_1, \dots, X_n)$ est un n-échantillon i.i.d. de loi de Bernoulli de paramètre p inconnu, alors

$$\left[\overline{X}_n - \frac{1}{2\sqrt{n\alpha}}, \overline{X}_n + \frac{1}{2\sqrt{n\alpha}}\right],$$

est un intervalle de confiance de p au niveau $1-\alpha$.

3.2 Exemple : estimation par intervalle de l'espérance d'une loi normale de variance donnée

Soit $(X_1, \dots, X_n)$ un n-échantillon i.i.d. de loi $\mathcal{N}(m, \sigma^2)$, où σ^2 est connu et m est le paramètre inconnu. On note Φ la fonction de répartition de la loi $\mathcal{N}(0,1)$.

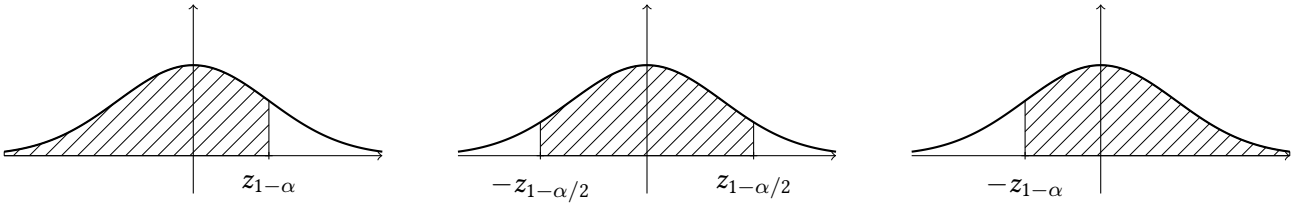
DÉFINITION 3.4 *Étant donné $\alpha \in]0,1[$, on appelle **quantile** de la loi $\mathcal{N}(0,1)$ à l'ordre α le réel $z_\alpha = \Phi^{-1}(\alpha)$.*

PROPOSITION 3.5 *Soient Z une variable aléatoire de loi $\mathcal{N}(0,1)$ et $\alpha \in]0,1[$.*
Chacun des intervalles

$$I_1 =]-\infty, z_{1-\alpha}], \quad I_2 = [-z_{1-\alpha/2}, z_{1-\alpha/2}] \quad \text{et} \quad I_3 = [-z_{1-\alpha}, +\infty[$$

est tel que $\mathbb{P}(Z \in I) = 1-\alpha$.

Remarque 3.6 L'intervalle I_2 est dit *bilatère*, les intervalles I_1 et I_3 sont dits *unilatères*.



La moyenne empirique

$$\overline{\mathrm{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{X}_i,$$

suit une loi normale, et l'on peut donc construire des intervalles de confiance de m. La proposition ci-dessous énonce un résultat précis ; le raisonnement doit être refait à chaque utilisation.

Proposition 3.7 *Si $z_{1-\alpha/2}$ désigne le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$, l'intervalle*

$$\left[\overline{\mathrm{X}}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{\mathrm{X}}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

est un intervalle de confiance de m au niveau de confiance $1-\alpha$.

Remarque 3.8 L'intervalle de confiance précédent est bilatère. On pourrait aussi proposer des intervalles de confiance unilatères.

3.3 Estimation par intervalle de confiance asymptotique

On a déjà remarqué que l'application de l'inégalité de Bienaymé-Tchebychev dans l'exemple 3.3 fournit un résultat qui peut être sensiblement précisé par le théorème limite central. Mais celui-ci donne seulement un résultat asymptotique, d'où la notion suivante.

Définition 3.9 *Soit $\alpha \in [0,1]$.*
*On appelle **intervalle de confiance asymptotique** de $g(\theta)$ au niveau de confiance $1-\alpha$ toute suite $([\mathrm{U}_n, \mathrm{V}_n])_{n\geqslant 1}$ vérifiant : pour tout $\theta \in \Theta$, il existe une suite (α_n) à valeurs dans $[0,1]$ et de limite α telle que*

$$\forall n \geqslant 1, \quad \mathbb{P}_\theta\big([\mathrm{U}_n \leqslant g(\theta) \leqslant \mathrm{V}_n]\big) \geqslant 1-\alpha_n.$$

Remarque 3.10 Par abus de langage, on dit aussi que $[\mathrm{U}_n, \mathrm{V}_n]$ est un intervalle de confiance asymptotique.

Exemple 3.11 (Estimation par intervalle du paramètre d'une loi de Bernoulli grâce au théorème limite central)
Si $(\mathrm{X}_1, \ldots, \mathrm{X}_n)$ est un n-échantillon i.i.d. de loi de Bernoulli de paramètre p inconnu, alors

$$\left[\overline{\mathrm{X}}_n - \frac{z_{1-\alpha/2}}{2\sqrt{n}}, \overline{\mathrm{X}}_n + \frac{z_{1-\alpha/2}}{2\sqrt{n}}\right]$$

est un intervalle de confiance asymptotique de p au niveau $1-\alpha$. En pratique, on considère qu'il s'agit d'un intervalle de confiance au niveau $1-\alpha$ dès que $np \geqslant 10$ et $n(1-p) \geqslant 10$.

Exemple 3.12 De retour sur l'exemple 1.1, on suppose que sur les $n = 2\,500$ personnes interrogées, 1 300 se sont prononcées pour A et 1 200 pour son adversaire. L'estimateur $\overline{\mathrm{X}}_n$ prend la valeur 0,52 mais peut-on raisonnablement affirmer pour autant que A sera élu ? Plus précisément, cette valeur est-elle significativement supérieure à 0,5 ?
L'intervalle de confiance à 95% pour θ fourni par la proposition précédente est alors $[0{,}50; 0{,}54]$. Avec l'inégalité de Bienaymé-Tchebychev, on obtiendrait $[0{,}47; 0{,}57]$.

Remarque 3.13 La méthode présentée dans l'exemple 3.11 dans le cas d'une loi de Bernoulli permet en fait d'estimer par intervalle de confiance l'espérance m_θ d'une loi admettant une variance σ_θ^2 si l'on connaît un réel $s > 0$ tel que $\sigma_\theta \leqslant s$ pour tout $\theta \in \Theta$.

Proposition 3.14 *On suppose que pour tout $\theta \in \Theta$, la loi ν_θ admet un moment d'ordre 4 et que son espérance m_θ ainsi que sa variance σ_θ^2 sont fonctions de θ.*

Dans ces conditions, l'écart-type empirique

$$S_n = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(X_k - \overline{X}_n\right)^2}$$

est un estimateur convergent de l'écart-type σ_θ.

La démonstration du résultat ci-dessous est à connaître et à savoir mettre en pratique.

THÉORÈME 3.15 *On suppose que pour tout $\theta \in \Theta$, la loi ν_θ admet un moment d'ordre 2 et que son espérance m_θ ainsi que sa variance $\sigma_\theta^2 > 0$ sont fonctions de θ.*
Dans ces conditions, si $z_{1-\alpha/2}$ désigne le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$ et $S_n > 0$ un estimateur convergent de l'écart-type σ_θ (par exemple l'écart-type empirique dans les conditions de la proposition 3.14), alors

$$\left[\overline{X}_n - z_{1-\alpha/2}\frac{S_n}{\sqrt{n}}, \overline{X}_n + z_{1-\alpha/2}\frac{S_n}{\sqrt{n}}\right]$$

est un intervalle de confiance asymptotique de m_θ de niveau $1 - \alpha$.

Remarque 3.16 Le théorème ci-dessus est encore valable sous l'hypothèse affaiblie $S_n \geqslant 0$.

Exemple 3.17 Si $(X_1, \ldots, X_n)$ est un n-échantillon i.i.d. de loi de Bernoulli de paramètre p inconnu, alors

$$\left[\overline{X}_n - z_{1-\alpha/2}\sqrt{\frac{\overline{X}_n\left(1-\overline{X}_n\right)}{n}}, \overline{X}_n + z_{1-\alpha/2}\sqrt{\frac{\overline{X}_n\left(1-\overline{X}_n\right)}{n}}\right]$$

est un intervalle de confiance asymptotique de p au niveau $1 - \alpha$.