Chapitre 16 : Estimation réelles

Introduction

0.1 Principe de la statistique inférentielle

Bien souvent, lorsqu'on étudie un phénomène aléatoire et une variable aléatoire X qui lui est liée, la loi de X n'est pas complètement spécifiée. Plus précisément, on connaît le type de loi de X (Bernoulli, géométrique, ...) mais pas son ou ses paramètres.

Exemple 1

On considère une urne ne contenant que des boules rouges et blanches. On tire et boule dans l'urne et on note X *la variable aléatoire valant* 1 *si la boule tirée est rouge et* 0 *sinon.*

1. *Quel type de loi suit la variable* X *?*

2. *Connaît-on son paramètre ? Pourquoi ?*

Le problème de l'estimation consiste alors à estimer la vraie valeur des paramètres à partir d'un échantillon de données obtenues en observant le phénomène se répéter : c'est ce qu'on appelle **la statistique inférentielle**.

Exemple 2

On considère l'urne de l'exemple précédent et on souhaite déterminer expérimentalement la proportion p de boules rouges. Pour cela :

(i) on tire ***avec remise*** *n boules dans l'urne ;*

(ii) on note la fréquence observée d'apparition des boules rouges :

$$\textit{fréquence observée} = \qquad .$$

Pourquoi cette fréquence observée permet-elle d'estimer le paramètre de loi suivie par X *?*

0.2 Modélisation mathématique

On considère une variable aléatoire X définie sur un espace probabilisable $(\Omega, \mathscr{A})$ dont la loi est à chercher parmi une famille de probabilités $(P_\theta)_{\theta\in\Theta}$ dépendant du paramètre θ ($\Theta \subset \mathbb{R}$).

Exemple 3

Dans les exemples précédents, on sait que X *suit une loi de Bernoulli.*

L'objectif est alors d'estimer la vraie valeur de θ ou parfois de $g(\theta)$ (où g est une fonction à valeurs réelles).

Dans toute la suite du chapitre, $(\Omega, \mathscr{A})$ **désigne un espace probabilisable muni d'une famille de probabilités** $(P_\theta)_{\theta\in\Theta}$ **dépendant d'un paramètre** θ **(**$\Theta \subset \mathbb{R}$**). La lettre** g **désignera un fonction définie sur** Θ **et à valeurs réelles.**

1 Estimation ponctuelle

1.1 Estimateur

Définition 1 (Échantillon)

Soient $n \in \mathbb{N}^*$ et X une variable aléatoire sur $(\Omega, \mathscr{A})$.

1. On appelle n**-échantillon de la loi de** X toute famille $(X_1, \dots, X_n)$ de variables aléatoires définies sur $(\Omega, \mathscr{A})$ mutuellement indépendantes et de même loi que X pour tout $\theta \in \Theta$.
2. Si $(X_1, \dots, X_n)$ est un n-échantillon, alors pour tout $\omega \in \Omega$ le n-uplet $(X_1(\omega), \dots, X_n(\omega))$ est appelé une **réalisation** de cet échantillon.

Exemple 4

On reprend l'exemple d'une urne ne contenant que des boules rouges et blanches. On tire une boule dans l'urne et on note X *la variable aléatoire valant* 1 *si la boule tirée est rouge et* 0 *sinon.*

1. *La variable* X *suit une loi de Bernoulli de paramètre inconnu.*
2. *On tire successivement* n*-fois avec remise une boule dans l'urne et on note* X_i *la variable aléatoire valant* 1 *si la boule tirée est rouge et* 0 *sinon.*
3. *Si deux personnes tirent tour à tour* 10 *boules dans l'urne, on obtient deux réalisations d'un* 10*-échantillon.*

Définition 2 (Estimateur)

Soit X une variable aléatoire réelle dont la loi dépend de θ.
Soient $n \in \mathbb{N}^*$ et $(X_1, \dots, X_n)$ un n-échantillon de la loi de X.

1. Si g est une fonction à valeurs réelles, on appelle **estimateur** de $g(\theta)$ toute variable aléatoire T_n de la forme :
$$T_n = \varphi(X_1, \dots, X_n)$$
où φ est une **fonction ne dépendant pas de** θ.
2. On appelle alors **estimation** de $g(\theta)$ toute réalisation $T_n(\omega)$ de T_n.

Exemple 5

Soit X *une variable aléatoire dont la loi dépend du paramètre* θ *et soit* $(X_1,\dots,X_n)$ *un* n*-échantillon de la loi de* X*. Alors les variables aléatoires suivantes sont des estimateurs de* $g(\theta)$ *:*

- $S_n = X_1 + \cdots + X_n$ *et* $\overline{X}_n = \frac{S_n}{n}$ *;*
- $X_1 \times \cdots \times X_n$ *;*
- $X_1 + \sqrt{X_n^2 + 1}$.

En revanche les variables aléatoires suivant n'en sont pas :

- $S_n - \theta$ *;*
- $\theta X_1 + \cdots + \theta^n X_n$.

Exemple important (Moyenne empirique)

Soit X une variable aléatoire et soit $(X_1,\dots,X_n)$ un n-échantillon de la loi de X.

La variable aléatoire $\overline{X}_n = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$ est un estimateur appelé **moyenne empirique** de l'échantillon.

Remarque 1

Le but d'un estimateur est de donner une estimation du paramètre θ*. Cependant la définition d'estimateur est très peu contraignante et beaucoup d'estimateurs parmi les exemples précédents sont peu pertinents.*

1. *Par exemple, si* X *suit une loi* $\mathscr{B}(p)$ *et que l'on veut estimer* p*, on a déjà vu (et on le reverra) qu'il est naturel d'utiliser l'estimateur* $\overline{X}_n$*. Au contraire, l'estimateur* $X_1 + \sqrt{X_n^2 + 1}$ *semble peu pertinent.*
2. *De même, si* X *suit une loi* $\mathscr{B}(p)$ *et que l'on veut estimer* $V(X) = p(1-p)$ *on peut considérer les estimateurs suivants :*
   (a) $\overline{X}_n(1-\overline{X}_n)$ *;*
   (b) $\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(X_k - \overline{X}_n)^2$.

Il convient donc de faire le tri parmi tous les estimateurs pour distinguer les "bons" des "mauvais". Pour cela, on va s'intéresser à l'écart entre l'estimateur et la valeur de $g(\theta)$.

1.2 Qualité d'un estimateur

1.2.1 Biais et biais asymptotique

Notation

Soit X une variable aléatoire sur $(\Omega, \mathscr{A})$ et soit $\theta \in \Theta$.
Si, considérée comme une variable aléatoire sur $(\Omega, \mathscr{A}, P_\theta)$, X possède une espérance alors on la note $E_\theta(X)$ pour signifier que l'espérance dépend du paramètre θ. De même, sous réserve d'existence, on notera $V_\theta(X)$ la variance de X considérée comme une variable aléatoire sur $(\Omega, \mathscr{A}, P_\theta)$.

Exemple 6

On lance une pièce et on considère la variable aléatoire X *qui vaut* 1 *si on obtient Pile et* 0 *sinon. On ne sait pas si la pièce est truquée.*

1. *La variable* X *suit une loi de Bernoulli de paramètre inconnu.*

2. *L'espérance et la variance de* X *dépendent du paramètre :*

Définition 3 (Biais)

Soit X une variable aléatoire réelle dont la loi dépend de θ.
Soient $n \in \mathbb{N}^*$ et T_n un estimateur de $g(\theta)$.

1. Si T_n possède une espérance pour tout $\theta \in \Theta$, on appelle **biais** de T_n le réel :
$$b_\theta(T_n) = E_\theta(T_n) - g(\theta).$$
2. L'estimateur T_n de $g(\theta)$ est dit **sans biais** si pour tout $\theta \in \Theta$, $b_\theta(T_n) = 0$.

Exemple important

Soit X une variable aléatoire possédant une espérance pour tout $\theta \in \Theta$ et soit $(X_1, \ldots, X_n)$ un n-échantillon de la loi de X.
La moyenne empirique $\overline{X}_n = \dfrac{1}{n}\sum_{k=1}^{n} X_k$ est un estimateur sans biais de $E_\theta(X)$.

Démonstration : Soit X une variable aléatoire possédant une espérance pour tout $\theta \in \Theta$ et soit $(X_1, \ldots, X_n)$ un n-échantillon de la loi de X.

■

Exemple 7

1. *Si* X *suit une loi de Bernoulli de paramètre* $\theta \in]0,1[$ *inconnu alors la moyenne empirique est un estimateur sans biais de* θ.
2. *Si* X *suit une loi de Poisson de paramètre* $\theta > 0$ *inconnu alors la moyenne empirique est un estimateur sans biais de* θ.

Test 1 (*Voir solution.*)

Soient $a \in \mathbb{R}_+^*$ *et* $n \in \mathbb{N}^*$. *On considère* $X_1, \ldots, X_n$ *une suite de variables aléatoires indépendantes de loi* $\mathcal{U}([0,a])$. *On pose* $M_n = \max(X_1, \ldots, X_n)$.

1. *Déterminer la fonction de répartition de* M_n. *En déduire que* M_n *est à densité et déterminer une densité.*
2. *Justifier que* M_n *est une estimateur du paramètre* a *et déterminer son biais.*

Test 2 (*Voir solution.*)

Soit X *une variable aléatoire possédant une espérance* m *et un moment d'ordre* 2 *noté* m_2. *Soit* $(X_1, \ldots, X_n)$ *un* n*-échantillon de la loi de* X. *On appelle* ***variance empirique*** *de l'échantillon la variable :*

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

1. *On considère* S_n^2 *comme un estimateur de* $V(X)$. *Déterminer son biais. Est-ce un estimateur sans biais ?*
2. *Montrer que* $\frac{n}{n-1}S_n^2$ *est un estimateur sans biais de* $V(X)$.

L'estimateur $\frac{n}{n-1}S_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_i - \overline{X}_n)^2$ *est appelée la variance empirique modifiée.*

Définition 4 (Estimateur asymptotiquement sans biais)

Soit X une variable aléatoire réelle dont la loi dépend de θ.
Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite d'estimateurs de $g(\theta)$ possèdant une espérance pour tout $\theta \in \Theta$.

1. On dit que la suite $(T_n)_{n\in\mathbb{N}^*}$ est **asymptotiquement sans biais** si pour tout $\theta \in \Theta$:

$$\lim_{n\to+\infty} b_\theta(T_n) = 0 \quad \text{ou encore} \quad \lim_{n\to+\infty} E_\theta(T_n) = g(\theta).$$

2. Par abus de langage, on dira souvent que T_n est asymptotiquement sans biais.

Exemple 8

Soit X une variable aléatoire suivant une loi de Bernoulli de paramètre $p \in]0,1[$ inconnu et soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires mutuellement indépendantes de même loi que X. On souhaite estimer la variance $p(1-p)$ de X à l'aide de l'estimateur $Y_n = \overline{X}_n(1-\overline{X}_n)$.

1. *Déterminons le biais de Y_n.*

2. *L'estimateur Y_n est-il sans biais ? Asymptotiquement sans biais ?*

Test 3 (***Voir solution.***)

On reprend l'énoncé du test 1. Montrer que M_n est une estimateur asymptotiquement sans biais du paramètre a.

Test 4 (*Voir solution.*)

*On reprend l'énoncé du test 2. Montrer que la **variance empirique** est un estimateur asymptotiquement sans biais de la variance.*

Remarque 2

Le biais mesure l'écart moyen entre les valeurs prises par l'estimateur et la quantité $g(\theta)$ à estimer. Si l'estimateur est sans biais, les valeurs de l'estimateur sont en moyenne très proches de $g(\theta)$. Cependant, à cause de phénomène de compensation, la moyenne des valeurs de l'estimateur peut être proche de $g(\theta)$ tout en ne prenant que des valeurs très éloignées de $g(\theta)$! Un tel estimateur ne fournira pas une estimation ponctuelle fiable.

Par exemple, on suppose qu'un observateur extérieur veut estimer la moyenne θ d'une classe au dernier contrôle de maths. Pour simplifier les choses, on va supposer que $\theta = 10$. Il décide de prendre l'estimateur T qui prend les valeurs 0 et 20 avec probabilité $\frac{1}{2}$.

1.2.2 Risque quadratique

Définition 5 (Risque quadratique)

Soit X une variable aléatoire réelle dont la loi dépend de θ.
Soient $n \in \mathbb{N}^*$ et T_n un estimateur de $g(\theta)$. Si pour tout $\theta \in \Theta$ la variable aléatoire T_n possède un moment d'ordre 2, on appelle **risque quadratique de** T_n le réel $r_\theta(T_n)$ défini par :

$$r_\theta(T_n) = E_\theta((T_n - g(\theta))^2).$$

Remarque 3

1. *Le risque quadratique mesure la moyenne des carrés des écarts entre l'estimateur et $g(\theta)$. Comme un carré est toujours positif, le phénomène de compensation évoqué dans la remarque 2 est éliminé.*

   Par exemple, pour l'estimateur de la remarque 2 le risque quadratique est :

2. *Un estimateur avec un risque quadratique faible prend des valeurs qui s'écartent peu par rapport à $g(\theta)$: il fournit donc de bonnes estimations.*

Théorème 1 (Décomposition biais-variance)

Soit X une variable aléatoire réelle dont la loi dépend de θ.
Soient $n \in \mathbb{N}^*$ et T_n un estimateur de $g(\theta)$ possédant un moment d'ordre 2 pour tout $\theta \in \Theta$.

1. On a alors : $r_\theta(T_n) = b_\theta(T_n)^2 + V_\theta(T_n)$.
2. En particulier, si T_n est sans biais alors : $r_\theta(T_n) = V_\theta(T_n)$.

Démonstration : Soit X une variable aléatoire réelle dont la loi dépend de θ.
Soient $n \in \mathbb{N}^*$ et T_n un estimateur de $g(\theta)$ possédant un moment d'ordre 2 pour tout $\theta \in \Theta$.

■

Exemple important

Soit X une variable aléatoire possédant une espérance m (paramètre à estimer) et une variance σ^2 et soit $(X_1,\dots,X_n)$ un n-échantillon de la loi de X.
La moyenne empirique $\overline{X}_n = \dfrac{1}{n}\sum_{k=1}^{n} X_k$ est un estimateur sans biais de m et

$$r_m(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Démonstration : Soit X une variable aléatoire possédant une espérance m et une variance σ^2 et soit $(X_1,\dots,X_n)$ un n-échantillon de la loi de X.

■

Test 5 (*Voir solution.*)

On reprend l'énoncé du test 1. Déterminer le risque quadratique de M_n *(en tant qu'estimateur du paramètre a).*

1.2.3 Estimateur convergent

Définition 6 (Estimateur convergent)

Soit X une variable aléatoire réelle dont la loi dépend de θ.
Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite d'estimateurs de $g(\theta)$.

1. On dit que la suite $(T_n)_{n\in\mathbb{N}^*}$ est **convergente** si pour tout $\theta\in\Theta$:

$$\forall \epsilon > 0, \quad \lim_{n\to+\infty} P_\theta(|T_n - g(\theta)| > \epsilon) = 0.$$

2. Par abus de langage, on dira souvent que T_n est convergent.

Remarque 4

Un estimateur convergent est un estimateur qui va donner, avec une forte probabilité, une bonne estimation de $g(\theta)$ pour des grands échantillons.

Exemple important

Soit X une variable aléatoire possédant une espérance et une variance pour tout $\theta \in \Theta$ et soit $(X_1, \ldots, X_n)$ un n-échantillon de la loi de X.
La moyenne empirique $\overline{X}_n = \dfrac{1}{n}\sum_{k=1}^{n} X_k$ est un estimateur convergent de $E_\theta(X)$.

Démonstration :

■

Théorème 2 (Condition suffisante de convergence)

Soit X une variable aléatoire réelle dont la loi dépend de θ.
Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite d'estimateurs de $g(\theta)$ possédant un moment d'ordre 2 pour tout $\theta \in \Theta$. Si, pour tout $\theta \in \Theta$ on a

$$\lim_{n\to+\infty} r_\theta(T_n) = 0,$$

alors $(T_n)_{n\in\mathbb{N}^*}$ est convergente.

Démonstration : Soit X une variable aléatoire réelle dont la loi dépend de θ.
Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite d'estimateurs de $g(\theta)$ possédant un moment d'ordre 2 pour tout $\theta \in \Theta$. On suppose que pour tout $\theta \in \Theta$ on a :

$$\lim_{n\to+\infty} r_\theta(T_n) = 0.$$

Exemple 9

Soit X *une variable aléatoire possédant une espérance* m *(paramètre à estimer) et une variance* σ^2 *et soit* $(X_1,\dots,X_n)$ *un* n*-échantillon de la loi de* X.
On a vu que la moyenne empirique $\overline{X}_n$ *possède un moment d'ordre* 2 *et que :*

$$r_m(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Test 6 (*Voir solution.*)

On reprend l'énoncé du test 1. L'estimateur M_n *du paramètre* a *est-il convergent ?*

2 Estimation par intervalle de confiance

Bien qu'il existe des critères pour juger la qualité d'un estimateur ponctuel (biais, risque quadratique, convergence), aucune certitude ne peut jamais être apportée quant au fait que l'estimation donne la valeur réelle du paramètre à estimer.
La démarche de l'estimation par intervalle de confiance consiste à trouver un intervalle aléatoire qui contienne $g(\theta)$ avec une probabilité minimale donnée.

2.1 Intervalle de confiance

Définition 7 (Intervalle de confiance)

Soit X une variable aléatoire réelle dont la loi dépend de θ.
Soient U_n et V_n deux d'estimateurs de $g(\theta)$ tels que pour tout $\theta \in \Theta$, $P_\theta(U_n \leqslant V_n) = 1$.
Soit $a \in [0,1]$.

1. On dit que l'intervalle $[U_n, V_n]$ est **un intervalle de confiance de** $g(\theta)$ **au niveau de confiance** $1-a$ si pour tout $\theta \in \Theta$ on a :
$$P_\theta([U_n \leqslant g(\theta) \leqslant V_n]) \geqslant 1-a.$$
2. Le réel a est alors appelé le **niveau de risque** de l'intervalle.

Remarque 5

1. *Dans beaucoup de cas, on choisit le niveau de risque* $a = 0.05$ *et on obtient un intervalle de niveau confiance* $1-a = 0.95$ *(c'est-à-dire à* 95%*).*
2. *L'inégalité*
$$P_\theta([U_n \leqslant g(\theta) \leqslant V_n]) \geqslant 1-a$$
signifie qu'avec un forte probabilité, la valeur $g(\theta)$ *à estimer est dans l'intervalle* $[U_n, V_n]$.
Par exemple, si on considère une urne ne contenant que des boules rouges et blanches et qu'on cherche à estimer la proportion θ *de boules rouges dans l'urne. Si l'on dispose de deux estimateurs tels que :*
$$P_\theta([U_n \leqslant g(\theta) \leqslant V_n]) \geqslant 0.95$$
alors cela signifie que si on réalise 100 *fois l'expérience aléatoire, dans au moins* 95 *des cas le paramètre* $g(\theta)$ *sera compris entre la valeur prise par* U_n *et la valeur prise par* V_n.
3. *Il est toutefois possible que les valeurs observées de* U_n *et* V_n *ne fournissent pas un encadrement de* $g(\theta)$ *: cela se produit avec un risque* a.

Méthode 1 (*Utilisation de l'inégalité de Bienaymé-Tchebychev*)

Soit X *une variable aléatoire réelle dont la loi dépend de* θ.
Soient $n \in \mathbb{N}^*$ *et* T_n *un estimateur de* $g(\theta)$ *sans biais et possédant un moment d'ordre* 2 *pour tout* $\theta \in \Theta$.

1. *Soit* $\epsilon > 0$. *D'après l'inégalité de Bienaymé-Tchebychev on a :*

2. *On sait que* T_n *est sans biais.*

3. *Supposons maintenant que l'on arrive à majorer* $V_\theta(T_n)$ *par une constante* v_n *indépendante de* θ.

Exemple 10

Soit X *une variable aléatoire réelle suivant une loi de Bernoulli de paramètre* p *à déterminer.*
Soient $n \in \mathbb{N}^*$ *et* $(X_1, \ldots, X_n)$ *un* n*-échantillon de la loi de* X. *On note* $\overline{X}_n$ *la moyenne empirique.*

1. *Quel est le biais de l'estimateur* $\overline{X}_n$ *? Possède-il un moment d'ordre* 2 *?*

2. *On applique l'inégalité de Bienaymé-Tchebychev à* $\overline{X}_n$.

3. *On en déduit un intervalle de confiance : soit* $\varepsilon > 0$.

Remarque 6

1. *L'intervalle que l'on obtient par cette méthode est centré en* $\overline{X}_n$ *et d'amplitude* 2ε. *Le réel* ε *s'appelle la* ***marge d'erreur*** *de l'intervalle.*
2. *A priori, on aimerait avoir la marge d'erreur la plus faible possible car cela donne un encadrement plus précis du paramètre. Cependant le niveau de confiance* $1 - \frac{1}{4n\varepsilon^2}$ *dépend de* ε *: ainsi plus la marge d'erreur est faible plus le niveau de confiance est faible.*
   En pratique, on cherche donc un bon équilibre entre précision de l'intervalle (c'est-à-dire amplitude ou marge d'erreur faible) et le niveau de confiance qu'on peut lui accorder (niveau de confiance élevé).

Exemple 11 (*Les sondages*)

Soit Ω *la population française. On veut déterminer la proportion* θ *d'individus qui aime les chats. On note* X *la variable aléatoire définie sur* $(\Omega, \mathscr{P}(\Omega))$ *par :*

$$\forall \omega \in \Omega, \quad X(\omega) = \begin{cases} 1 & \text{si l'individu } \omega \text{ aime les chats} \\ 0 & \text{sinon.} \end{cases}$$

Ainsi X *suit la loi de Bernoulli de paramètre* θ *(à estimer).*
Comme il est difficile d'interroger tous les français, on va seulement interroger un échantillon de n *français pris au hasard dans la population : on considère donc un échantillon* $(X_1, \ldots, X_n)$ *un* n*-échantillon de la loi de* X.

1. *Proposer un estimateur sans biais de* θ.
2. *Pour tout* $\varepsilon > 0$ *déterminer un intervalle de confiance d'amplitude* 2ε *et donner son niveau de confiance.*
3. *On se propose d'interroger* 3000 *personnes (c'est-à-dire qu'on prend* $n = 3000$*).*
   (a) *Le tableau suivant donne l'amplitude en fonction du niveau de confiance souhaité.*

	Marge d'erreur ε *(en %) en fonction du niveau de confiance* $1-a$ *(en %) avec* $n = 3000$ *fixé*							
$1-a$	70	75	80	85	90	95	97	99
ε	1.7	1.8	2	2.4	2.9	4.1	5.3	9.1

(b) Le tableau suivant donne le niveau de confiance en fonction de l'amplitude souhaitée.

	Niveau de confiance $1-a$ *(en %) en fonction de la marge d'erreur* ε *(en %) avec* $n = 3000$ *fixé*							
ε	1.5	2	2.5	3	3.5	4	4.5	5
$1-a$	63	79	87	91	93	95	96	97

(c) Que peut-on remarquer?

(d) En pratique, on se fixe plutôt une marge d'erreur et un niveau de confiance et on adapte la taille de l'échantillon n (le nombre de personnes sondées).

	Nombre de sondés n en fonction de la marge d'erreur ε *(en %) et du niveau de confiance* $1-a$ *(en %)*							
ε \ $1-a$	70	75	80	85	90	95	97	99
0.5	33 333	40 000	50 000	66 667	100 000	200 000	333 333	1 000 000
1	8 333	10 000	12 500	16 667	25 000	50 000	83 333	250 000
1.5	3 704	4 444	5 556	7 407	11 111	22 222	37 037	111 111
2	2 083	2 500	3 125	4 167	6 250	12 500	20 083	62 500
2.5	1 333	1 600	2 000	2 667	4 000	8 000	13 333	40 000
3	926	1 111	1 389	1 852	2 778	5 556	9 259	27 778
3.5	680	816	1 020	1 361	2 041	4 082	6 803	20 408
4	521	625	781	1 042	1 563	3 125	5 208	15 625

2.2 Intervalle de confiance asymptotique

Définition 8 (Intervalle de confiance asymptotique)

Soit X une variable aléatoire réelle dont la loi dépend de θ.
Soient $(U_n)_{n\in\mathbb{N}*}$ et $(V_n)_{n\in\mathbb{N}*}$ deux suites d'estimateurs de $g(\theta)$ tels que pour tout $\theta\in\Theta$ et pour tout $n\geqslant 1$, $P_\theta(U_n\leqslant V_n)=1$.
Soit $a\in[0,1]$.

1. On dit que l'intervalle $[U_n,V_n]$ est **un intervalle de confiance asymptotique de** $g(\theta)$ **au niveau de confiance** $1-a$ si pour tout $\theta\in\Theta$ il existe une suite de réels $(a_n)_{n\in\mathbb{N}*}$ à valeurs dans $[0,1]$, de limite a telle que :
$$\forall n\in\mathbb{N}^*,\quad P_\theta([U_n\leqslant g(\theta)\leqslant V_n])\geqslant 1-a_n.$$
2. Le réel a est alors appelé le **niveau de risque** de l'intervalle.
3. Par abus de langage on dit aussi que $[U_n,V_n]$ est un intervalle de confiance asymptotique.

Méthode 2 (*Utilisation du théorème central limite*)

Soit X *une variable aléatoire réelle d'espérance* θ *inconnue et de variance* σ^2 ***non nulle connue*** *indépendante de* θ.
On considère $(X_n)_{n\in\mathbb{N}*}$ *indépendantes et de même la loi que* X *et on note* $\overline{X}_n$ *la moyenne empirique.*
Soit $a\in[0,1]$ *et cherchons un intervalle de confiance asymptotique de niveau* $1-a$.

1. *Rappeler l'espérance et la variance de* $\overline{X}_n$ *:*
2. *Appliquer le théorème central limite.*
3. *Soit* $\epsilon>0$. *On détermine un intervalle de confiance asymptotique.*

4. *Conclusion : en posant $t_a = \phi^{-1}\left(1-\frac{a}{2}\right)$, où ϕ est la fonction de répartition d'une variable aléatoire de loi normale centrée réduite, on obtient l'intervalle asymptotique suivant :*

Le réel t_a est appelé ***le quantile d'ordre*** $1-\frac{a}{2}$ *de la loi $\mathcal{N}(0,1)$.*

Remarque 7

Les intervalles de confiance asymptotiques ne donnent de garanties qu'asymptotiquement.
Cependant, le théorème central limite donne de bonnes approximations même pour des valeurs de n faibles (dès que $n \geqslant 30$).

Exemple 12

Soit X une variable aléatoire réelle suivant une loi de Bernoulli de paramètre p à déterminer.
Soient $n \in \mathbb{N}^*$, $(X_1, \ldots, X_n)$ un n-échantillon de X et $a \in [0,1]$.

1. *Que donne le TCL dans ce cas ?*
2. *⚠ Ici, les bornes de l'intervalle dépendent du paramètre à estimer (car la variance dépend de p). Ce n'est donc pas un intervalle de confiance. On va utiliser la majoration $p(1-p) \leqslant \frac{1}{4}$ pour obtenir un vrai intervalle de confiance asymptotique.*

Remarque 8

L'intervalle que l'on obtient par cette méthode est centré en $\overline{X}_n$ et d'amplitude $\frac{t_a}{\sqrt{n}}$.

Exemple 13 (*Les sondages*)

On reprend l'exemple du sondage mais cette fois en utilisant l'intervalle de confiance asymptotique obtenu avec le TCL. La marge d'erreur ε vaut $\frac{t_a}{2\sqrt{n}}$.
Le tableau suivant donne le nombre n de personnes à sonder en fonction de l'amplitude et du niveau de confiance souhaités.

	Nombre de sondés n en fonction de la marge d'erreur ε *(en %) et du niveau de confiance* $1-a$ *(en %)*							
ε \ $1-a$	70	75	80	85	90	95	97	99
0.5	10 816	13 689	16 384	20 736	26 896	38 416	44 089	66 049
1	2 704	3 422	4 096	5 184	6 724	9 604	11 722	16 512
1.5	1 202	1 521	1 820	2 304	2 988	4 268	5 232	7 339
2	676	852	1 024	1 296	1 681	2 401	2 943	4 128
2.5	433	548	655	829	1076	1537	1 884	2 642
3	300	380	455	576	747	1067	1 308	1 835
3.5	221	279	334	423	549	784	961	1 348
4	169	214	256	324	420	600	736	1 032
Valeur de t_a	1.04	1.17	1.28	1.44	1.64	1.96	2.17	2.57

3 Objectifs

1. Comprendre les notions d'estimation, d'estimateur.
2. Savoir déterminer le biais d'un estimateur.
3. Savoir déterminer le risque quadratique d'un estimateur.
4. Savoir montrer qu'un estimateur est sans biais et convergent.
5. Savoir déterminer un intervalle de confiance avec l'inégalité de Bienaymé-Tchebychev.
6. Savoir déterminer un intervalle de confiance avec le théorème central limite.