Estimation

I. Introduction : Des données à une modélisation

On se demande quelle proportion de la population française est climatosceptique (pour adapter les programmes scolaires par exemple, les campagnes de sensibilisation, les réformes imposées aux citoyens et aux entreprises, etc).
Pour cela, on demande à plusieurs personnes leur opinion sur ce sujet.
On interroge n personnes dans un panel de français et on inscrit les résultats dans un tableau. On code le climatoscepticisme par 1, et par 0 l'opinion inverse.

Individu numéro	1	2	3	4	5	6	7	8	9	10	11	12
Résultats	0	0	1	0	1	1	1	0	0	0	0	1

Les sondés sont numérotés et leur choix est noté x_i.
Par exemple, le $3^{\text{ème}}$ individu interrogé est climatosceptique : $x_3 = 1$.

Première tentative d'exploitation des résultats du tableau

- On peut commencer par décrire les résultats obtenus.
  Sur la série statistique précédente (tableau de données), on peut :
  - × calculer la moyenne empirique.

$$\overline{x_{12}} = \frac{\sum_{i=1}^{12} x_i}{12} = \frac{5}{12} \approx 0,42$$

  - × calculer l'écart-type empirique.

$$s_{12} = \sqrt{\frac{1}{12} \sum_{i=1}^{12} \left(x_i - \overline{x_{12}}\right)^2} \approx 0,47$$

- C'est le monde de la statistique **descriptive** : on décrit les données collectées. Cette étude est limitée. Savoir que 7 personnes sur notre échantillon de 12 personnes ne sont pas climatosceptiques ne permet pas de répondre à la question : « la population française est-elle majoritairement climatosceptique » ?
- Ce cours porte sur la théorie de l'estimation : on cherche à déduire des renseignements sur une population à partir de la connaissance d'un échantillon. On parle dans ce cas de **statistique inférentielle**.

Modélisation statistique de l'expérience

- Le mathématicien se place à l'instant précédant les interrogations des individus. Ces interrogations sont alors considérées comme des expériences aléatoires.
- On poursuit en notant X_1, X_2, …, X_{12}, … les v.a.r. définies par :

$$\forall i \in \mathbb{N}^*, \quad X_i \ : \ \Omega \ \to \ \mathbb{R}$$

$$\omega \ \mapsto \ \begin{cases} 1 & \text{si le } i^{\text{ème}} \text{ individu interrogé est climato-sceptique} \\ 0 & \text{sinon} \end{cases}$$

- Les résultats x_i sont alors les réalisations des variables aléatoires X_i, on les appelle les valeurs observées ou les données.
  On note la différence entre l'utilisation des majuscules pour les variables aléatoires et les minuscules pour leurs réalisations (les données).
- Le v.a.r. X_i sont supposées :
  - × indépendantes.
    Pour comprendre ce point, il faut penser à la population comme une immense urne : chaque boule est un individu. On choisit succesivement n individus et on procède avec remise (ce qui permet d'assurer l'indépendance). Agir de la sorte peut faire craindre d'interroger plusieurs fois le même individu. C'est possible mais peu probable si le nombre n d'individus du panel est faible par rapport à la taille de la population.
  - × de même loi.
    Ici, la loi commune est une loi de Bernoulli de paramètre **inconnu** noté θ. Le paramètre θ est alors la probabilité du succès : « être climatosceptique ». C'est ce paramètre θ que l'on cherche à estimer.

  Dans la suite, on dira que $(X_1, \ldots, X_n)$ est un n-échantillon de la v.a.r. X où $X \hookrightarrow \mathcal{B}(\theta)$. Par ailleurs, la v.a.r. :

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

  sera appelée **moyenne empirique**.

Remarque

- Il est à noter que l'aléa ne se situe pas dans ce que va répondre un sondé donné : ce choix est entièrement déterministe puisque dicté par ses goûts / expériences. L'aléa se situe en réalité uniquement dans le fait que ce sondé ait été choisi pour faire partie ou non du panel.
  L'utilisation de v.a.r. correspond donc à une vision de l'esprit, un choix de la modélisation.
- On peut tomber sur des panels standards : la moyenne empirique de l'échantillon correspond à la moyenne théorique θ (celle de la population). Mais on peut aussi tomber sur des panels atypiques : la moyenne empirique est très écartée de la moyenne théorique. Devant ce constat, deux solutions s'offrent à nous :
  - × s'entraîner à avoir la main chaude. Lors du tirage au sort des participants, on fait en sorte de tomber sur un panel dont la proportion de climatosceptiques correspond à la proportion de la population. Cela a peu de chance d'aboutir.
  - × assumer que le résultat issu des données puisse ne pas correspondre au résultat théorique. Cela ne signifie pas que le résultat est inutilisable mais simplement qu'il y a un risque qu'il ne corresponde pas au résultat théorique. Tout l'enjeu est alors de savoir quel risque on prend en considérant que le résultat issu des données est proche du résultat théorique.
- On conçoit assez aisément que la taille du panel joue un rôle important : la probabilité de tomber sur un cas atypique a tendance à décroître lorsque la taille de l'échantillon augmente.

- On connaît donc la forme de la loi commune des observations (une loi de Bernoulli), mais on ignore son paramètre. L'objet de la statistique est de dire des choses sur ce paramètre.
  Par exemple, se demander si la majorité de la population française est climatosceptique revient à se demander si :
  $$\theta > \frac{1}{2} \quad \text{ou} \quad \theta < \frac{1}{2}$$
  (très peu probable en France... on s'intéresserait plutôt à une question du type : $\theta < 10\%$)
- Si ce panel est vraiment très grand, les enquêteurs n'ont pas le temps d'interroger tout le monde et ne peuvent accéder à la vraie valeur de θ, d'où la nécessité de méthodes statistiques pour, malgré tout, pouvoir dire des choses avec un degré de confiance raisonnable sur θ.
- En résumé, le but du chapitre est d'étudier comment une connaissance d'un échantillon (calcul de la moyenne empirique) nous permet d'obtenir une information sur θ (moyenne théorique). Comme on l'a vu, on ne peut espérer que le résultat des données corresponde de manière sûre au résultat théorique. On peut toutefois concevoir obtenir certains garanties. En particulier, on formalisera le type de déclaration suivante :

  > « le résultat théorique θ est ε-proche du résultat issu des données avec une forte probabilité »

II. Estimation ponctuelle

On considère un phénomène aléatoire et on s'intéresse à une variable aléatoire X qui lui est liée, dont on suppose que la loi de probabilité n'est pas complètement spécifiée. On se restreint au cas où la forme de la loi est connue à un (ou deux) paramètres près que l'on cherche à estimer.
Dans les cas précédent, on cherchait à estimer le paramètre θ d'une loi de Bernoulli. Toutes les lois de Bernoulli sont au départ envisageables : on considère initialement la famille contenant toutes les lois $\mathcal{B}(\theta)$ pour $\theta \in \Theta = [0, 1]$.

II.1. Notion de n-échantillon

Définition (n-échantillon)
Soit X une v.a.r. définie sur un espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$.
On appelle n-échantillon de la v.a.r. X tout n-uplet $(X_1, \ldots, X_n)$ de v.a.r. , définies sur $(\Omega, \mathscr{A}, \mathbb{P})$:

× indépendantes,

× de même loi que X.

Commentaire

Lorsqu'on réalise n fois une expérience E, on obtient un n-échantillon de données.
On peut donc définir les v.a.r. $X_1, \ldots, X_n$ de la façon suivante : lorsqu'on réalise n fois l'expérience E, X_i correspond à la valeur prise par X lors de la i-ème réalisation de E.

II.2. Estimateur et estimation

II.2.a) Définition

Définition *(Estimateur)*

Soit X une v.a.r. dont la loi dépend d'un paramètre θ.

Soit $n \in \mathbb{N}^*$ et soit $(X_1, \ldots, X_n)$ un n-échantillon de la v.a.r. X.

- On appelle **estimateur de** θ toute v.a.r. T_n, qui s'exprime en fonction des v.a.r. $(X_1, \ldots, X_n)$ et dont l'expression ne fait pas mention du paramètre θ.

- Autrement dit, T_n est un estimateur de θ s'il existe une fonction φ de n variables telle que :

$$T_n = \varphi(X_1, \ldots, X_n)$$

- Si φ permet de définir un estimateur de θ alors, on appelle **estimation de** θ tout réel de la forme :

$$\varphi(x_1, \ldots, x_n)$$

où $(x_1, \ldots, x_n) \in X_1(\Omega) \times \ldots \times X_n(\Omega)$.

- Lorsque l'estimateur T_n possède une espérance (resp. une variance), on la note $\mathbb{E}_\theta(T_n)$ (resp. $\mathbb{V}_\theta(T_n)$) *(lire espérance / variance sous θ).*

Remarque

Le fait de noter $\mathbb{E}_\theta(T_n)$ l'espérance de T_n rappelle que celle-ci dépend a priori du paramètre θ. Il en va évidemment de même de la loi de probabilité de la variable T_n : un événement (par exemple le succès dans une épreuve de Bernoulli) prend des probabilités différentes selon la valeur de θ. En toute rigueur, il convient donc de l'indiquer en notant $\mathbb{P}_\theta$ l'application probabilité en jeu, car elle dépend à son tour de la valeur du paramètre estimé. Un estimateur T_n a donc une dépendance en θ (issu de la dépendance en θ de la loi de X). Pour autant, cela n'autorise pas à faire apparaître θ dans l'expression de T_n (c'est une contrainte différente).

Exemple *(d'estimateurs)*

Soit X une v.a.r. dont la loi dépend d'un paramètre θ.

Soit $n \in \mathbb{N}^*$ et soit $(X_1, \ldots, X_n)$ un n-échantillon de la v.a.r. X.

- Les v.a.r. $S_n = X_1 + \ldots + X_n$ et $\overline{X_n}$ sont des estimateurs de θ.
  (attention, l'expression de la v.a.r. $\overline{X}_n^*$ peut dépendre de θ !)
- La v.a.r. $X_1 \times \ldots \times X_n$ est un estimateur de θ.
- Si $(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n$, la v.a.r. $\sum_{i=1}^{n} \lambda_i \cdot X_i$ est un estimateur de θ.
- Les v.a.r. X_1 et $X_2 - \sqrt{X_1 + X_n}$ sont des estimateurs de θ.

Par contre :

× la v.a.r. constante égale à θ n'est pas un estimateur de θ.

× la v.a.r. $X_1 + \ldots + X_n - \theta$ n'est pas un estimateur de θ.

En effet, ces v.a.r. font apparaître θ dans leur expression ce qui est interdit dans la définition.

Commentaire

- Dans la pratique, c'est donc la valeur prise par T_n après réalisation de l'expérience aléatoire qui sert à estimer le paramètre ; cette estimation ne dépend que de l'échantillon $(x_1, x_2, ..., x_n)$ observé. On remarquera par ailleurs que la définition est très générale : on ne dit rien sur la qualité de l'estimation.
- Dans certains cas, on ne cherche pas à estimer le paramètre θ, mais son image $g(\theta)$ par une certaine fonction : $g(\theta)$ peut par exemple être l'espérance ou la variance de la variable aléatoire X estimée. Cela ne change rien à la définition précédente, où il suffit de remplacer θ par $g(\theta)$.

  Par exemple, pour estimer la variance $p(1-p)$ de la variable de Bernoulli X du paragraphe **I.2**, il est logique de considérer l'estimateur $\overline{X}_n(1 - \overline{X}_n)$.

⚠ Le paramètre θ **ne doit pas intervenir** dans la définition de l'estimateur T_n, sinon ce n'est pas un estimateur !

II.2.b) Un exemple classique d'estimateur

Théorème 1.

Soit X une v.a.r. dont la loi dépend d'un paramètre θ.

Soit $(X_1, \ldots, X_n)$ un n-échantillon de la v.a.r. X.

- *Alors la v.a.r.* $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ *est un estimateur de θ.*
- *Cet estimateur est appelée* **moyenne empirique** *de l'échantillon $(X_1, \ldots, X_n)$.*

Remarque

- Le but d'un **estimateur** $T_n = \varphi(X_1, \ldots, X_n)$ est de fournir une **estimation** $\varphi(x_1, \ldots, x_n)$ du paramètre inconnu θ que l'on cherche à estimer. Ainsi, même si la définition d'estimateur est peu contraignante, cela ne signifie pas pour autant que tous les estimateurs listés précédemment sont pertinents. Il convient alors de faire le tri parmi tous ces estimateurs. Une idée simple est alors de signifier qu'un bon estimateur est un estimateur qui fournit de bonnes estimations. Il s'agit alors de mesurer l'écart entre les valeurs que peut prendre l'estimateur T_n et la valeur du paramètre θ. Dans ce cours, on aborde deux mesures de la qualité d'un estimateur : le biais et le risque quadratique.
- Notons par ailleurs que dans certains cas, on ne cherche pas à estimer le paramètre θ, mais plutôt son image $g(\theta)$ par une certaine fonction g. Au passage, en fonction de ce que signifie le paramètre θ par rapport à la loi de X, on proposera naturellement des estimateurs différents.
  Par exemple, si $X \hookrightarrow \mathcal{B}(p)$ alors :
  - × pour estimer p, il paraît naturel de proposer l'estimateur moyenne empirique $\overline{X_n}$ puisque $\mathbb{E}(X) = p$.
  - × pour estimer $\mathbb{V}(X) = p(1-p)$, on peut proposer deux estimateurs :
    1) $\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X_n}\right)^2$, estimateur appelé variance empirique (estimateur naturel pour trouver une estimation d'une variance, quelle que soit la loi de X),
    2) $\overline{X_n}\left(1 - \overline{X_n}\right)$, estimateur qui tire parti de l'expresion de $\mathbb{E}(X)$.

II.3. Qualité de l'estimateur

II.3.a) Biais d'un estimateur

Définition

Soit X une v.a.r. dont la loi dépend d'un paramètre θ.

Soit $n \in \mathbb{N}^*$ et soit T_n un estimateur de θ.

On suppose que T_n admet une espérance.

- Si T_n admet une espérance, on appelle **biais de l'estimateur** T_n le réel :

$$\boxed{b_\theta(T_n) = \mathbb{E}_\theta\big(T_n - \theta\big) = \mathbb{E}_\theta(T_n) - \theta}$$

- Si le paramètre estimé est $g(\theta)$, l'expression du biais devient :

$$b_\theta(T_n) = \mathbb{E}_\theta\big(T_n - g(\theta)\big) = \mathbb{E}_\theta(T_n) - g(\theta)$$

- On dit que T_n est un **estimateur sans biais de** θ lorsque $b_\theta(T_n) = 0$, ou encore : $\mathbb{E}_\theta(T_n) = \theta$. Dans le cas contraire, on parlera d'estimateur biaisé.

II.3.b) Estimateur asymptotiquement sans biais

Définition

Soit X unve v.a.r. dont la loi dépend d'un paramètre θ.

Soit $(T_n)_{n\in\mathbb{N}^*}$ une d'estimateurs de θ (resp. $g(\theta)$).

On suppose que pour tout $n \in \mathbb{N}^*$, T_n admet une espérance.

- On dit que la suite d'estimateurs $(T_n)_{n\in\mathbb{N}^*}$ est **asymptotiquement sans biais** lorsque :

$$\lim_{n\to+\infty} b_\theta(T_n) = 0 \quad \text{ou encore} \quad \lim_{n\to+\infty} \mathbb{E}_\theta(T_n) = \theta$$

$$(\text{resp. } \lim_{n\to+\infty} \mathbb{E}_\theta(T_n) = g(\theta))$$

- On dit aussi, par abus de langage, que l'estimateur T_n est asymptotiquement sans biais.

Remarque

- Le biais mesure l'écart **moyen** entre les valeurs prises par l'estimateur et le paramètre θ à estimer. Si l'estimateur est sans biais, les valeurs de l'estimateur sont **en moyenne** très proches de θ. Intuitivement, un estimateur sans biais peut donc sembler de meilleure qualité qu'un estimateur biaisé. En réalité, l'absence de biais ne permet pas de conclure quant à la qualité de l'estimateur : il existe des estimateurs sans biais de piètre qualité.
- Imaginons que l'on souhaite déterminer une estimation d'un paramètre θ. Pour des raisons pédagogiques, on considère que ce paramètre est connu et nul ($\theta = 0$). On propose un estimateur T de θ dont la loi est la suivante :
  - $\times$ $T(\Omega) = \{-5, 5\}$.
  - $\times$ $\mathbb{P}\big([T = -5]\big) = \frac{1}{2} = \mathbb{P}\big([T = 5]\big)$.

  La v.a.r. T est finie. Elle admet donc une espérance. De plus :

$$\begin{aligned}\mathbb{E}_\theta(T-\theta) &= \mathbb{E}_\theta(T-0) \;=\; \mathbb{E}_\theta(T)\\ &= -5\times\mathbb{P}\big([T=-5]\big) + 5\times\mathbb{P}\big([T=5]\big) \;=\; -\tfrac{5}{2}+\tfrac{5}{2} \;=\; 0\end{aligned}$$

  Ainsi, T est un estimateur sans biais de θ.
  Rappelons que le but d'un estimateur est de fournir une estimation. Par définition, cette estimation est l'une des valeurs de l'estimateur : cela sera donc soit 5, soit -5. L'une comme l'autre sont des mauvaises estimations de $\theta = 0$ (on peut remplacer -1 par -10^6 et 1 par 10^6 si les valeurs -1 et 1 semblent trop proches de 0).
  Ici, l'estimateur s'écarte beaucoup de la valeur θ à estimer. Pour autant, un phénomène de compensation se produit : un fort écart négatif (-5) et compensé par un fort écart positif (5), de sorte à obtenir une espérance de T qui est nulle, ce qui fait, ici, que T est une estimateur sans biais.
- Pour éviter ce phénomène de compensation, on peut penser à considérer tout écart de manière positive. Cela reviendrait à introduire la mesure :

$$\mathbb{E}\big(|T_n - \theta|\big)$$

  Si cette mesure moyenne est assurément digne d'intérêt, elle n'est pas forcément simple à déterminer. Une autre possibilité est de considérer les écarts quadratiques. C'est ce que l'on considère dans le chapitre suivant.

II.3.c) Risque quadratique

Définition

Soit X une v.a.r. dont la loi dépend d'un paramètre θ (resp. $g(\theta)$).

Soit $n \in \mathbb{N}^*$ et soit T_n un estimateur de θ.

On suppose que T_n admet un moment d'ordre 2.

On appelle **risque quadratique de l'estimateur** T_n le réel :

$$\boxed{r_\theta(T_n) \;=\; \mathbb{E}_\theta\left((T_n-\theta)^2\right)} \quad (\text{resp. } \mathbb{E}_\theta\left((T_n-g(\theta))^2\right))$$

Remarque

- Le risque quadratique mesure la moyenne des carrés des écarts au paramètre θ (on remarquera la similitude avec la définition de la variance). Comme un carré est toujours positif, une telle mesure produit une accumulation des écarts et élimine le phénomène de compensation.
  Par exemple, pour l'exemple précédent on obtient :

$$\begin{aligned}\mathbb{E}_\theta\big((T-\theta)^2\big) &= \mathbb{E}_\theta\big((T-0)^2\big) \;=\; \mathbb{E}_\theta\big(T^2\big)\\ &= (-5)^2\times\mathbb{P}\big(\,[T=-5]\,\big)+5^2\times\mathbb{P}\big(\,[T=5]\,\big)\\ &= \tfrac{25}{2}+\tfrac{25}{2} \;=\; 25\end{aligned}$$

- Si un estimateur possède un risque quadratique faible, alors les valeurs qu'il prend s'écartent peu de θ. Les valeurs de cet estimateur fourniront donc, avec une probabilité forte, de bonnes estimations de θ.

Entre deux estimateurs de θ, on préférera toujours celui dont le risque quadratique est le plus faible.

Théorème 2. *(Décomposition biais - variance)*

Soit X une v.a.r. dont la loi dépend d'un paramètre θ (resp. $g(\theta)$).

Soit $n \in \mathbb{N}^*$ et soit T_n un estimateur de θ.

On suppose que T_n admet un moment d'ordre 2.

1) *On a alors :*

$$\boxed{r_\theta(T_n) \;=\; \mathbb{V}_\theta(T_n)+\big(b_\theta(T_n)\big)^2}$$

2) *Si on suppose de plus que T_n est sans biais alors :*

$$\boxed{r_\theta(T_n) \;=\; \mathbb{V}_\theta(T_n)}$$

Démonstration.

- Comme la v.a.r. T_n admet un moment d'ordre 2, alors elle admet un risque quadratique.
- De plus :

$$\begin{aligned}
& \mathbb{V}_\theta(T_n) + \big(b_\theta(T_n)\big)^2 \\
= \; & \mathbb{E}_\theta\big((T_n)^2\big) - \big(\mathbb{E}_\theta(T_n)\big)^2 + \big(\mathbb{E}_\theta(T_n) - \theta\big)^2 && \textit{(par la formule de Kœnïg-Huygens)} \\
= \; & \mathbb{E}_\theta\big((T_n)^2\big) - \cancel{\big(\mathbb{E}_\theta(T_n)\big)^2} + \cancel{\big(\mathbb{E}_\theta(T_n)\big)^2} - 2\,\theta\,\mathbb{E}_\theta(T_n) + \theta^2 \\
= \; & \mathbb{E}_\theta\big(T_n^2 - 2\,\theta\,T_n + \theta^2\big) && \textit{(par linéarité de l'espérance)} \\
= \; & \mathbb{E}_\theta\big((T_n - \theta)^2\big) \\
= \; & r_\theta(T_n)
\end{aligned}$$

□

Commentaire

Explication compromis biais - variance

Exercice 1

On reprend l'exercice **??** et on note $(X_1, \ldots, X_n)$ un n-échantillon de X.

- Calculer le biais et le risque quadratique de l'estimateur $T_n = X_1$.
- Calculer le biais et le risque quadratique de l'estimateur $\bar{X}_n$.
- Quel estimateur de θ choisir ?

II.4. Estimateur convergent

II.4.a) Définition

Définition

Soit X une v.a.r. dont la loi dépend d'un paramètre θ.

Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite d'estimateurs de θ (resp. $g(\theta)$).

- On dit que la suite d'estimateurs $(T_n)_{n\in\mathbb{N}^*}$ de θ (resp. $g(\theta)$) est convergente si :

$$\boxed{\forall \varepsilon > 0, \; \lim_{n\to+\infty} \mathbb{P}\big(\,[\,|T_n - \theta\,| > \varepsilon\,]\,\big) = 0}$$

$$\Big(\text{resp. } \forall \varepsilon > 0, \; \lim_{n\to+\infty} \mathbb{P}\big(\,[\,|T_n - g(\theta)\,| > \varepsilon\,]\,\big) = 0\Big)$$

On dit aussi, par abus de langage, que l'estimateur T_n est convergent.

- On peut réécrire la propriété précédente :

$$\boxed{\text{L'estimateur } T_n \text{ est convergent} \quad \Leftrightarrow \quad T_n \xrightarrow[n\to+\infty]{\mathbb{P}} \theta}$$

Ainsi, un estimateur convergent est un estimateur qui converge vers le paramètre à estimer.

II.4.b) Démontrer en pratique qu'un estimateur est convergent

Théorème 3.

Soit X une v.a.r. dont la loi dépend d'un paramètre θ (resp. $g(\theta)$).

Soit $n \in \mathbb{N}^*$ et soit T_n un estimateur de θ.

On suppose que T_n admet un moment d'ordre 2.

$$\boxed{\lim_{n\to+\infty} r_\theta(T_n) = 0 \quad \Rightarrow \quad T_n \text{ est un estimateur convergent}}$$

Démonstration.
Soit $\varepsilon > 0$. La v.a.r. $(T_n - \theta)^2$:

× admet une espérance,

× est à valeurs positives.

Ainsi, par inégalité de Markov (avec $a = \varepsilon^2$) :

$$\mathbb{P}_\theta\left(\left[(T_n-\theta)^2 > \varepsilon^2\right]\right) \leqslant \frac{\mathbb{E}_\theta((T_n-\theta)^2)}{\varepsilon^2}$$

Or, par stricte croissance de $x \mapsto \sqrt{x}$ sur $[0, +\infty[$:

$$\mathbb{P}_\theta\left(\left[(T_n-\theta)^2 > \varepsilon^2\right]\right) = \mathbb{P}_\theta\big([|T_n-\theta| > \varepsilon]\big)$$

On en déduit :

$$0 \leqslant \mathbb{P}_\theta\big([|T_n-\theta| > \varepsilon]\big) \leqslant \frac{\mathbb{E}_\theta((T_n-\theta)^2)}{\varepsilon^2} = \frac{r_\theta(T_n)}{\varepsilon^2}$$

Or :

× $\lim\limits_{n\to+\infty} 0 = 0$,

× $\lim\limits_{n\to+\infty} \dfrac{r_\theta(T_n)}{\varepsilon^2} = 0$.

Ainsi, par théorème d'encadrement : $\lim\limits_{n\to+\infty} \mathbb{P}_\theta([|T_n-\theta| > \varepsilon]) = 0$. □

Commentaire

La propriété annoncée par ce théorème est très naturelle. On suppose que la limite du risque quadratique est nulle. Autrement dit, asymptotiquement parlant (en $+\infty$), l'estimateur T_n a un écart quadratique nul avec le paramètre θ à estimer. On en déduit que T_n est asymptotiquement égal à θ. Cela s'énonce rigoureusement par le fait que T_n converge (en probabilité) vers θ.

II.5. Estimation de l'espérance

Proposition 1.

Soit X une v.a.r. admettant une espérance m et une variance.

Soit $(X_1, \ldots, X_n)$ un n-échantillon de la v.a.r. X.

Alors la moyenne empirique $\overline{X}_n = \dfrac{X_1 + \cdots + X_n}{n}$ est un estimateur sans biais convergent de m.

Preuve.

- Montrons que $\bar{X}_n$ est un estimateur sans biais de m.
  - × La v.a.r. $\bar{X}_n$ admet une espérance en tant que combinaison linéaire de v.a.r. qui en admettent une.

× De plus :

$$\mathbb{E}(\bar{X}_n) \;=\; \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \;=\; \frac{1}{n}\sum_{i=1}^{n} m = m$$

Donc $\bar{X}_n$ est un estimateur sans biais de m.

- Montrons que $\bar{X}_n$ est un estimateur convergent de m.
  × La v.a.r. $\bar{X}_n$ admet une variance en tant que combinaison linéaire de v.a.r. qui en admettent une.
  × Comme $\bar{X}_n$ est un estimateur sans biais de m, alors $r_m(\bar{X}_n) = \mathbb{V}(\bar{X}_n)$. Or, comme les X_i sont indépendants :

$$\mathbb{V}(\bar{X}_n) \;=\; \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \;=\; \frac{1}{n^2}\,\mathbb{V}\left(\sum_{i=1}^{n} X_i\right) \;=\; \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i) \;=\; \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \;=\; \frac{\sigma^2}{n} \underset{n\to+\infty}{\longrightarrow} 0$$

Donc, d'après la proposition précédente, $\bar{X}_n$ est un estimateur convergent de m. □

Exercice 2
Proposer un estimateur du paramètre λ d'une loi de Poisson et donner son risque quadratique.

III. Intervalle de confiance

III.1. Exemple du sondage

- Revenons sur le premier exemple du climatoscepticisme. Il illustre le cas du sondage. Rappelons la formalisation :
  × $X \hookrightarrow \mathcal{B}(\theta)$ où θ est la proportion de climatosceptiques de la population française en entier.
  × $(X_1, \ldots, X_n)$ est un n-échantillon de la v.a.r. X.
  × $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$, moyenne empirique du n-échantillon est utilisé comme estimateur de l'espérance de X. C'est un bon estimateur puisqu'il est convergent.
- Le rôle d'un estimateur est de fournir une estimation. Dans cet exemple, on a pris comme valeur de θ une valeur possible de $\overline{X_{12}}$. Il apparaît peu probable que cette valeur soit une bonne approximation de θ : il convient certainement d'augmenter la taille de l'échantillon considéré.
- On a annoncé en début de chapitre que l'on formaliserait la garantie d'approximation suivante :

  « le résultat théorique θ est ε-proche du résultat issu des données avec une forte probabilité »

  C'est le rôle des intervalles de confiance que de formaliser cette déclaration.

III.2. Intervalle de confiance (exact)

Définition (Intervalle de confiance)
Soit X une v.a.r. dont la loi dépend d'un paramètre θ (à estimer).
Soient $(U_n)_{n\in\mathbb{N}^*}$ et $(V_n)_{n\in\mathbb{N}^*}$ deux suites d'estimateurs de θ (resp. $g(\theta)$).
On suppose de plus que : $\mathbb{P}_\theta\big([U_n \leqslant V_n]\big) = 1$.
Soit $\alpha \in]0,1[$.

- On dit que $[U_n, V_n]$ est un intervalle de confiance de θ (resp. $g(\theta)$) au niveau de confiance $1-\alpha$ si :

$$\mathbb{P}_\theta\Big([U_n \leqslant \theta \leqslant V_n]\Big) \;\geqslant\; 1-\alpha$$

$$\mathbb{P}_\theta([\theta \in [U_n, V_n]])$$

$$\Big(\text{resp. } \mathbb{P}_\theta\Big([U_n \leqslant g(\theta) \leqslant V_n]\Big) \;\geqslant\; 1-\alpha\Big)$$

- Le réel α est appelé le niveau de risque de l'intervalle.

Commentaire

- Il est assez classique de choisir un niveau de risque $\alpha = 0.05$. On obtient alors un intervalle au niveau de confiance $1 - \alpha = 0.95$ (95%).
- Interrogeons-nous sur la signification de la minoration définissant un intervalle de confiance :
$$\mathbb{P}_\theta\Big([U_n \leqslant \theta \leqslant V_n] \Big) \geqslant 0.95 \quad (*)$$
Cela signifie qu'avec une forte probabilité, le paramètre à estimer se trouve dans l'intervalle aléatoire $[U_n, V_n]$. Les v.a.r. U_n et V_n sont des estimateurs de θ. Elles ont pour but de fournir une estimation de θ. En pratique, ce sont donc les valeurs observées de U_n et V_n qui nous intéressent. L'inégalité $(*)$ permet de considérer que si l'on effectue 100 fois l'expérience aléatoire (le sondage dans le cas de l'exemple initial) alors, dans au moins 95 cas, on aura : $\theta \in [u_n, v_n]$ où u_n (resp. v_n) est la valeur prise par U_n (resp. v_n) au cours de l'expérience.
- On retiendra qu'il est toujours possible que le résultat obtenu lors de l'expérience (l'intervalle $[u_n, v_n]$ censé contenir θ) ne soit pas un bon encadrement de θ. Plus précisément, cela se produit avec un risque α. Il est évident que l'on souhaite que ce risque soit faible et le choix $\alpha = 0.05$ semble raisonnable.

MÉTHODO

Utilisation de l'inégalité de Bienaymé-Tchebychev pour obtenir un intervalle de confiance d'un paramètre θ à l'aide d'un estimateur sans biais de θ

Soit X une v.a.r. dont la loi dépend d'un paramètre θ (à estimer).

Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite d'estimateurs de θ.

On suppose que pour tout $n \in \mathbb{N}^*$: $\mathbb{E}_\theta(T_n) = \theta$.
(T_n est un estimateur sans biais de θ)

1) Soit $\varepsilon > 0$. D'après l'inégalité de Bienaymé-Tchebychev :
$$\begin{array}{rcl} \mathbb{P}_\theta\big([\, |T_n - \mathbb{E}_\theta(T_n)| > \varepsilon \,] \big) & \leqslant & \dfrac{\mathbb{V}_\theta(T_n)}{\varepsilon^2} \\ \shortparallel & & \\ \mathbb{P}_\theta\big([\, |T_n - \theta| > \varepsilon \,] \big) & & \end{array}$$

2) Supposons qu'on arrive à majorer $\mathbb{V}_\theta(T_n)$ indépendamment de θ.
Autrement dit, supposons qu'il existe une suite réelle (v_n) telle que : $\forall n \in \mathbb{N}^*$, $\mathbb{V}_\theta(T_n) \leqslant v_n$. Alors on obtient :
$$\mathbb{P}_\theta\big([\, |T_n - \theta| > \varepsilon \,] \big) \leqslant \frac{v_n}{\varepsilon^2}$$

3) On en déduit alors :
$$\begin{array}{rcl} 1 - \mathbb{P}_\theta\big([\, |T_n - \theta| > \varepsilon \,] \big) & \geqslant & 1 - \dfrac{v_n}{\varepsilon^2} \\ \shortparallel & & \\ \mathbb{P}_\theta\big([\, |T_n - \theta| \leqslant \varepsilon \,] \big) & & \\ \shortparallel & & \\ \mathbb{P}_\theta\big([\, T_n - \varepsilon \leqslant \theta \leqslant T_n + \varepsilon \,] \big) & & \end{array}$$

On en conclut que $[T_n - \varepsilon, T_n + \varepsilon]$ est un intervalle de confiance de θ au niveau de confiance $1 - \dfrac{v_n}{\varepsilon^2}$.

Illustration classique : estimation de l'espérance pour une loi de Bernoulli

Soit X une v.a.r. de loi de Bernoulli $\mathcal{B}(p)$ où p est un paramètre à estimer.

Soit $n \in \mathbb{N}^*$ et soit $(X_1, \dots, X_n)$ un n-échantillon de la v.a.r. X.

Notons $\overline{X_n} = \dfrac{1}{n} \sum\limits_{i=1}^{n} X_i$.

Rappelons que la v.a.r. $\overline{X_n}$ admet une espérance et une variance. De plus :

× $\mathbb{E}(\overline{X_n}) \;=\; \mathbb{E}\left(\dfrac{1}{n} \sum\limits_{i=1}^{n} X_i\right) \;=\; \dfrac{1}{n} \sum\limits_{i=1}^{n} \mathbb{E}(X_i) \;=\; \dfrac{1}{n} \sum\limits_{i=1}^{n} p = p.$

× $\mathbb{V}(\overline{X_n}) \;=\; \mathbb{V}\left(\dfrac{1}{n} \sum\limits_{i=1}^{n} X_i\right) \;=\; \dfrac{1}{n^2} \sum\limits_{i=1}^{n} \mathbb{V}(X_i) = \dfrac{1}{n^2} \sum\limits_{i=1}^{n} p(1-p) \;=\; \dfrac{p(1-p)}{n}.$

1) Soit $\varepsilon > 0$. D'après l'inégalité de Bienaymé-Tchebychev :

$$\mathbb{P}\big(\,[\;|\overline{X_n} - p| \geqslant \varepsilon\;]\,\big) \;\leqslant\; \frac{p(1-p)}{n\,\varepsilon^2}$$

2) À l'aide de la majoration classique $p\,(1-p) \leqslant \dfrac{1}{4}$, on obtient :

$$\mathbb{P}\big(\,[\;|\overline{X_n} - p| > \varepsilon]\,\big) \;\leqslant\; \frac{p(1-p)}{n\,\varepsilon^2} \;\leqslant\; \frac{1}{4\,n\,\varepsilon^2}$$

3) On en déduit :

$$\begin{array}{ccc} 1 - \mathbb{P}\big(\,[\;|\overline{X_n} - p\;| > \varepsilon\;]\,\big) & \geqslant & 1 - \dfrac{1}{4\,n\,\varepsilon^2} \\ \shortparallel & & \\ \mathbb{P}\big(\,[\;|\overline{X_n} - p\;| \leqslant \varepsilon\;]\,\big) & & \\ \shortparallel & & \\ \mathbb{P}\big(\,[\;\overline{X_n} - \varepsilon \leqslant p \leqslant \overline{X_n} + \varepsilon\;]\,\big) & & \end{array}$$

En posant : $U_n = \overline{X_n} - \varepsilon$, $V_n = \overline{X_n} + \varepsilon$ et $\alpha = \dfrac{1}{4\,n\,\varepsilon^2}$ on en conclut :

$$\boxed{\mathbb{P}\big(\,[\;U_n \leqslant p \leqslant V_n\;]\,\big) \;\geqslant\; 1 - \alpha}$$

Ainsi, $[\overline{X_n} - \varepsilon, \overline{X_n} + \varepsilon]$ est un intervalle de confiance du paramètre p au niveau de confiance $1 - \dfrac{1}{4\,n\,\varepsilon^2}$.

Remarque

Il y a plusieurs remarques à faire sur l'intervalle obtenu :

× cet intervalle est centré en $\overline{X_n}$.

× l'amplitude de cet intervalle est : $V_n - U_n = (\overline{X_n} + \varepsilon) - (\overline{X_n} - \varepsilon) = 2\,\varepsilon$.
Il est à noter que le réel ε a été choisi en début de démonstration avec pour seule contrainte : $\varepsilon > 0$.
Ce réel est appelé **marge d'erreur** de l'intervalle. Il est possible, avec cette méthode de produire des intervalles avec marge d'erreur fixée à l'avance.
A priori, une marge d'erreur faible est préférable.

× la marge d'erreur de l'intervalle influe directement sur le niveau de risque car : $\alpha = \frac{1}{4\,n\,\varepsilon^2}$. Ainsi, une marge d'erreur faible produit un niveau de risque élevé et un niveau de confiance faible.

Rappelons que le but d'un estimateur est de fournir une estimation. Lorsque l'on procède au sondage, on détermine une valeur $\overline{x_n}$ prise par l'estimateur $\overline{X_n}$. On peut alors considérer, grâce à l'analyse précédente, que le paramètre p à estimer se trouve dans l'intervalle $[\overline{x_n} - \varepsilon, \overline{x_n} + \varepsilon]$ avec niveau de confiance $1 - \frac{1}{4\,n\,\varepsilon^2}$. Il s'agit alors de trouver un bon équilibre entre la précision de l'intervalle (mesure de l'amplitude ou de la marge d'erreur) et le niveau de confiance qu'on peut lui accorder (mesure du risque ou du niveau de confiance).

Équilibre marge d'erreur / niveau de confiance

Explicitons le lien entre marge d'erreur et niveau de confiance.
Dans la suite, on fixe $n = 2500$ ce qui correspond à un sondage auprès de 2500 personnes.

- Pour $n = 2500$ fixé et α donné, on obtient les valeurs suivantes pour la marge d'erreur $\varepsilon = \dfrac{1}{\sqrt{4\,n\,\alpha}}$.

	Calcul de la marge d'erreur ε (en %) en fonction du niveau de confiance $1-\alpha$ (en %) avec $n = 2500$ fixé							
$1-\alpha$	70	75	80	85	90	95	97,5	99
ε	1,8	2	2,2	2,6	3,2	4,5	6,3	10

La précision se dégrade lorsque le niveau de confiance augmente.

- Pour $n = 2500$ fixé et ε donné, on obtient les valeurs suivantes pour le niveau de confiance $1 - \alpha = 1 - \dfrac{1}{4\,n\,\varepsilon^2}$.

	Calcul du niveau de confiance $1-\alpha$ (en %) en fonction de la marge d'erreur ε (en %) avec $n = 2500$ fixé							
ε	0,5	1	1,5	2	2,5	3	3,5	4
$1-\alpha$		0	56	75	84	89	92	94

Le niveau de confiance augmente lorsque l'on dégrade la précision.

À RETENIR

- Améliorer la précision (diminuer la marge d'erreur ε) de l'intervalle, c'est augmenter le risque et ainsi diminuer le niveau de confiance.
- Dégrader la précision (augmenter la marge d'erreur ε) de l'intervalle, c'est diminuer le risque et ainsi augmenter le niveau de confiance.

Le point de vue des instituts de sondage

Lorsqu'un sondage est effectué, il faut systématiquement se poser la question des garanties aléatoires de précision sur lesquelles il se fonde. Pour les instituts de sondage, la question est donc de savoir combien de personnes il faut interroger pour obtenir un niveau de confiance $(1-\alpha)$ élevé et une précision importante (marge d'erreur ε faible).

Dans le tableau suivant, on calcule $n = \dfrac{1}{4\,\alpha\,\varepsilon^2}$ pour différentes valeurs du couple (ε, α).

	Nombre de sondés en fonction de la marge d'erreur ε (en %) et du niveau de confiance $1-\alpha$ (en %) souhaités							
ε \ $1-\alpha$	70	75	80	85	90	95	97,5	99
0,5	33333	40000	50000	66667	100000	200000	400000	1000000
1	8333	10000	12500	16667	25000	50000	100000	250000
1,5	3704	4444	5556	7407	11111	22222	44444	111111
2	2083	2500	3125	4167	6250	12500	25000	62500
2,5	1333	1600	2000	2667	4000	8000	16000	40000
3	926	1111	1389	1852	2778	5556	11111	27778
3,5	680	816	1020	1361	2041	4082	8163	20408
4	521	625	781	1042	1563	3125	6250	15625

- Comme mentionné précédemment, le niveau de confiance $1-\alpha = 0,95$ est assez classique. Avec un tel niveau de confiance, on considère qu'il y a 95% de chances de tomber sur un panel standard. Lorsque c'est le cas, le paramètre réel se retrouve dans l'intervalle $[\overline{x_n} - \varepsilon, \overline{x_n} + \varepsilon]$.
  - × On peut souhaiter obtenir un résultat très précis. Pour un sondage concernant des élections, savoir qu'un candidat est évalué à 19% plus ou moins 0,5% serait idéal. Du point de vue du sondeur, cela voudrait dire interroger $n = 200000$ personnes. C'est inenvisageable pour des raisons évidentes de coût.
  - × L'institut de sondage doit alors revoir ses objectifs à la baisse. En interrogeant $n = 3125$ personnes, il assure avec une probabilité de 95% qu'un candidat est évalué à 19% plus ou moins 4%. Le coût est tout à fait envisageable mais le résultat semble alors un peu trop imprécis.
- Les résultats de ce dernier tableau démontrent que la méthode permettant d'obtenir un intervalle de confiance par inégalité de Bienaymé-Tchebychev est peu exploitable lorsque l'on cherche à obtenir des résultats relativement précis. Cela provient du fait que l'inégalité de Bienaymé-Tchebychev qui s'applique à toute v.a.r. (sans exploitation de la loi de celle-ci) est assez peu précise. Les instituts de sondage se basent sur une autre méthode consistant à obtenir un intervalle de confiance à l'aide du théorème central limite. Ce théorème énonce un résultat de convergence en loi. On sait que cette convergence se fait rapidement (des valeurs faibles de n fournissent de très bonnes approximations du résultat). En conséquence, on peut espérer obtenir des garanties aléatoires de précision fortes avec un nombre de sondés plus faible. C'est l'objet du paragraphe suivant.

Exercice 3

On considère un n-échantillon $(X_1, \dots, X_n)$ d'une loi de Bernoulli de paramètre θ et $\bar{X}_n$ la moyenne empirique associée à l'échantillon. On rappelle que $\bar{X}_n$ est un estimateur sans biais de θ.

1. En utilisant l'inégalité de Bienaymé-Tchebychev, montrer, pour tout $\varepsilon > 0$:

$$\mathbb{P}_\theta\big(\big[\bar{X}_n - \varepsilon \leqslant \theta \leqslant \bar{X}_n + \varepsilon\big]\big) \geqslant 1 - \frac{1}{4n\varepsilon^2}$$

2. En déduire un intervalle de confiance de θ au niveau de confiance $0,95$.
On trouve : $\left[\bar{X}_n - \dfrac{1}{\sqrt{0,2n}}\ ,\ \bar{X}_n + \dfrac{1}{\sqrt{0,2n}}\right]$.

3. On réalise un sondage auprès de 500 personnes. 280 affirment vouloir voter pour M. A aux prochaines élections, soit 56% des sondés. Peut-on affirmer, avec un risque d'erreur de 5% que M. A sera élu aux prochaines élections ?

Commentaire

- Notez que l'aléa (le fait qu'on ne soit sûr qu'à 95% et non 100% de cet intervalle théorique) est causé par les événements atypiques, comme par exemple, obtenir au moins 90 fois la réponse « je suis climatosceptique » sur 100 individus interrogés, alors que ce pourcentage serait par ailleurs égal à 10% dans la population totale.
  Cet événement a un probabilité faible mais non nulle, et quand on somme les probabilités de tous les événements atypiques, que l'on exclut, on arrive à 5%.
- En grossissant l'intervalle, on peut en proposer un qui soit valide, à 99% par exemple ; en déterminant ε tel que $\dfrac{1}{4n\varepsilon^2} = \dfrac{1}{100}$ et en reprenant les calculs de l'exercice III.2. Enfin, pour être sûr à 100% de son assertion, il est nécessaire de proposer l'intervalle $[0, 1]$, mais c'est ridicule : c'est comme si l'expérience statistique n'avait pas servie.

- S'il y a effectivement une probabilité de 95% au moins pour que θ appartienne à l'intervalle aléatoire construit sur les X_i, cette probabilité est de 0 ou 1 concernant l'intervalle réalisé $[11\ ,\ 55]$. En effet, θ appartient ou pas à un tel intervalle déterministe, il n'y a pas d'autre issue. On pourrait déterminer ce fait en interrogeant tous les membres du panel (mais cela prendrait du temps). En pratique, on ne le fait pas et on espère avoir eu de la chance lors de la réalisation de l'intervalle de confiance.
  Pour éclairer cette distinction, pensez à un tirage du Loto : étant donnée une grille, avant le tirage, il lui est associé une (faible) probabilité de gain du gros lot ; après le tirage, la grille est soit gagnante, soit perdante. Évidemment, à parier, on parierait plutôt sur le fait de ne pas avoir eu le gros lot.
- Ici, l'indication de la probabilité *a priori* de 95% dans le cas des intervalles de confiance nous rassure quant à la validité de ce que l'on obtient après l'expérience sans toutefois que ce sentiment rassurant puisse être une certitude mathématique. On ajuste une indication de bonne confiance statistique.

III.3. Intervalle de confiance asymptotique

Définition (Intervalle de confiance asymptotique)

Soit X une v.a.r. dont la loi dépend d'un paramètre θ (à estimer).

Soient $(U_n)_{n\in\mathbb{N}^*}$ et $(V_n)_{n\in\mathbb{N}^*}$ deux suites d'estimateurs de θ (resp. $g(\theta)$).

On suppose de plus : $\forall n \in \mathbb{N}^*$, $\mathbb{P}_\theta\big([U_n \leqslant V_n]\big) = 1$.

Soit $\alpha \in [0,1]$.

- On dit que $[U_n, V_n]$ est un intervalle de confiance asymptotique de θ (resp. $g(\theta)$) au niveau de confiance $1-\alpha$ si :

$$\lim_{n\to+\infty} \quad \mathbb{P}_\theta\Big([U_n \leqslant \theta \leqslant V_n]\Big) \quad \geqslant \quad 1-\alpha$$

$$\mathbb{P}_\theta([\theta \in [U_n, V_n]])$$

$$\Big(\text{resp. } \mathbb{P}_\theta\Big([U_n \leqslant g(\theta) \leqslant V_n]\Big) \geqslant 1-\alpha\Big)$$

- Le réel $\alpha \in [0,1]$ est appelé le niveau de risque de l'intervalle.

MÉTHODO **Utilisation du TCL**

Soit X une v.a.r. . On suppose que X :

× admet une espérance m **inconnue**, qu'on cherche à estimer.

× admet une variance σ^2 **connue** et non nulle.

Soit $(X_1, \ldots, X_n)$ un n-échantillon de la v.a.r. X.

Soit $\alpha \in [0,1]$ (on cherche un intervalle de confiance de m au niveau de confiance $1-\alpha$).

Enfin, on note : $\overline{X_n} = \dfrac{X_1 + \cdot + X_n}{n}$.

Rappelons que la v.a.r. $\overline{X_n}$ admet une espérance et une variance. De plus :

$$\mathbb{E}(\overline{X_n}) = m \quad \text{et} \quad \mathbb{V}(\overline{X_n}) = \frac{\sigma^2}{n}$$

1) Alors les v.a.r. X_1, …, X_n sont :

× indépendantes,

× de même loi,

× admettent une variance non nulle.

Ainsi, par théorème central limite :

$$\overline{X_n}^* = \frac{\overline{X_n} - m}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[n\to+\infty]{\mathscr{L}} Z$$

où $Z \hookrightarrow \mathcal{N}(0,1)$.

2) Soit $x \in [0, +\infty[$. On a de plus :

$$\begin{aligned}
\mathbb{P}\left(\left[-x \leqslant \overline{X_n}^* \leqslant x\right]\right) &= \mathbb{P}\left(\left[-x \leqslant \sqrt{n}\,\frac{\overline{X_n} - m}{\sigma} \leqslant x\right]\right) \\
&= \mathbb{P}\left(\left[-\frac{\sigma\,x}{\sqrt{n}} \leqslant \overline{X_n} - m \leqslant \frac{\sigma\,x}{\sqrt{n}}\right]\right) \\
&= \mathbb{P}\left(\left[-\overline{X_n} - \frac{\sigma\,x}{\sqrt{n}} \leqslant -m \leqslant -\overline{X_n} + \frac{\sigma\,x}{\sqrt{n}}\right]\right) \\
&= \mathbb{P}\left(\left[\overline{X_n} + \frac{\sigma\,x}{\sqrt{n}} \geqslant m \geqslant \overline{X_n} - \frac{\sigma\,x}{\sqrt{n}}\right]\right)
\end{aligned}$$

On obtient ainsi :

$$\begin{aligned}\mathbb{P}\left(\left[\overline{X_n} - \frac{\sigma\, x}{\sqrt{n}} \leqslant m \leqslant \overline{X_n} + \frac{\sigma\, x}{\sqrt{n}}\right]\right) &= \mathbb{P}\left(\left[-x \leqslant \overline{X_n}^{*} \leqslant x\right]\right) \\ &\underset{n\to+\infty}{\longrightarrow} \mathbb{P}([-x \leqslant Z \leqslant x]) \\ &= \Phi(x) - \Phi(-x) \\ &= \Phi(x) - (1 - \Phi(x)) \;=\; 2\,\Phi(x) - 1\end{aligned}$$

3) On cherche alors une valeur x telle que : $2\,\Phi(x) - 1 \geqslant 1 - \alpha$. Or :

$$\begin{aligned}2\,\Phi(x) - 1 \geqslant 1 - \alpha &\Leftrightarrow 2\,\Phi(x) \geqslant 2 - \alpha \\ &\Leftrightarrow \Phi(x) \geqslant 1 - \tfrac{\alpha}{2} \\ &\Leftrightarrow x \geqslant \Phi^{-1}\big(1 - \tfrac{\alpha}{2}\big)\end{aligned}$$

On note alors : $q_{1-\frac{\alpha}{2}} = \Phi^{-1}\big(1 - \frac{\alpha}{2}\big)$.
Ce nombre est appelé **quantile d'ordre** $1 - \frac{\alpha}{2}$.
On le notera par la suite t_α pour plus de lisibilité.

4) D'après ce qui précède, on a :

$$\lim_{n\to+\infty} \mathbb{P}\left(\left[\overline{X_n} - \frac{\sigma}{\sqrt{n}}\, t_\alpha \leqslant m \leqslant \overline{X_n} + \frac{\sigma}{\sqrt{n}}\, t_\alpha\right]\right) \;=\; 2\Phi(t_\alpha) - 1 \;\geqslant\; 1 - \alpha$$

Ainsi, $\left[\overline{X_n} - \dfrac{\sigma}{\sqrt{n}}\, t_\alpha \;,\; \overline{X_n} + \dfrac{\sigma}{\sqrt{n}}\, t_\alpha\right]$ est un intervalle de confiance asymptotique de m au niveau de confiance $1 - \alpha$.

Il est à noter que ce type d'intervalle n'offre qu'une garantie asymptotique (en $+\infty$). Toutefois, on sait que le théorème central limite offre de très bonnes approximations pour des valeurs de n faibles (dès que $n \geqslant 30$).
Comme la taille du panel de sondés dépasse largement 30 personnes, on considère que la limite a lieu.

Illustration classique : estimation de l'espérance pour une loi de Bernoulli

Reprenons l'exemple précédent où $X \hookrightarrow \mathcal{B}(p)$ où p est le paramètre à estimer.

- En reprenant l'étude précédente, on obtient :

$$\lim_{n\to+\infty} \mathbb{P}\left(\left[\overline{X_n} - \frac{\sqrt{p(1-p)}}{\sqrt{n}}\, t_\alpha \leqslant p \leqslant \overline{X_n} + \frac{\sqrt{p(1-p)}}{\sqrt{n}}\, t_\alpha\right]\right) \geqslant 1 - \alpha$$

Attention : les extrémités de l'encadrement dépendent du paramètre à estimer p. On utilise donc la majoration classique $p(1-p) \leqslant \dfrac{1}{4}$ pour obtenir des estimateurs de p.

- Comme $p(1-p) \leqslant \dfrac{1}{4}$, par croissance de $x \mapsto \sqrt{x}$ sur $[0, +\infty[$:

$$\sqrt{p(1-p)} \leqslant \frac{1}{2} \quad \text{et} \quad -\sqrt{p(1-p)} \geqslant -\frac{1}{2}$$

On en déduit :

$$\left[\overline{X_n} - \frac{\sqrt{p(1-p)}}{\sqrt{n}}\, t_\alpha \leqslant p \leqslant \overline{X_n} + \frac{\sqrt{p(1-p)}}{\sqrt{n}}\, t_\alpha\right]$$

$$\subset \quad \left[\overline{X_n} - \frac{1}{2\sqrt{n}}\, t_\alpha \leqslant p \leqslant \overline{X_n} + \frac{1}{2\sqrt{n}}\, t_\alpha\right]$$

Ainsi, par croissance de $\mathbb{P}$:

$$\mathbb{P}\left(\left[\overline{X_n} - \frac{\sqrt{p(1-p)}}{\sqrt{n}}\, t_\alpha \leqslant p \leqslant \overline{X_n} + \frac{\sqrt{p(1-p)}}{\sqrt{n}}\, t_\alpha\right]\right)$$

$$\leqslant \quad \mathbb{P}\left(\left[\overline{X_n} - \frac{1}{2\sqrt{n}}\, t_\alpha \leqslant p \leqslant \overline{X_n} + \frac{1}{2\sqrt{n}}\, t_\alpha\right]\right)$$

On en conclut, par transitivité :

$$\lim_{n\to+\infty} \mathbb{P}\left(\left[\overline{X_n} - \frac{1}{2\sqrt{n}}\, t_\alpha \leqslant p \leqslant \overline{X_n} + \frac{1}{2\sqrt{n}}\, t_\alpha\right]\right) \geqslant 1-\alpha$$

Ainsi, $\left[\overline{X_n} - \dfrac{1}{2\sqrt{n}}\, t_\alpha \ ,\ \overline{X_n} + \dfrac{1}{2\sqrt{n}}\, t_\alpha\right]$ est un intervalle de confiance asymptotique de p au niveau de confiance $1-\alpha$.

Commentaire

On peut faire le même genre de remarques sur l'intervalle obtenu :
× cet intervalle est centré en $\overline{X_n}$.
× l'amplitude de cet intervalle est :

$$\left(\overline{X_n} + \frac{1}{2\sqrt{n}}\ t_\alpha\right) - \left(\overline{X_n} - \frac{1}{2\sqrt{n}}\ t_\alpha\right) = \frac{t_\alpha}{\sqrt{n}}$$

Équilibre marge d'erreur / niveau de confiance

Rappelons les liens entre marge d'erreur ε et niveau de confiance $1-\alpha$:

$$\varepsilon = \frac{t_\alpha}{2\sqrt{n}} \quad \text{et} \quad 1-\alpha = 2\,\Phi(t_\alpha) - 1$$

Lorsque le niveau de confiance $1-\alpha$ augmente, $2\,\Phi(t_\alpha)-1$ augmente, ce qui démontre que t_α augmente. Dans ce cas, la marge d'erreur ε augmente elle aussi. Cet intervalle de confiance asymptotique possède les mêmes propriétés que l'intervalle de confiance exact (obtenu par l'inégalité de BT).

À RETENIR

- Améliorer la précision (diminuer la marge d'erreur ε) de l'intervalle, c'est augmenter le risque et ainsi diminuer le niveau de confiance.
- Dégrader la précision (augmenter la marge d'erreur ε) de l'intervalle, c'est diminuer le risque et ainsi augmenter le niveau de confiance.

Le point de vue des instituts de sondage

	Nombre de sondés en fonction de la marge d'erreur ε (en %) et du niveau de confiance $1-\alpha$ (en %) souhaités							
ε \ $1-\alpha$	70 (1,04)	75 (1,17)	80 (1,28)	85 (1,44)	90 (1,64)	95 (1,96)	97,5 (2,26)	99 (2,57)
0,5	10816	13689	16384	20736	26896	38416	51076	66049
1	2704	3422	4096	5184	6724	9604	12769	16512
1,5	1202	1521	1820	2304	2988	4268	5674	7339
2	676	856	1024	1296	1681	2401	3192	4128
2,5	433	548	655	829	1076	1537	2043	2642
3	300	380	455	576	747	1067	1419	1835
3,5	221	279	334	423	549	784	1042	1348
4	169	214	256	324	420	600	798	1032

- Sur la première ligne, on a placé entre parenthèse la valeur de t_α correspondante au niveau de confiance $1-\alpha$ considéré.
  Par exemple, si $1-\alpha = 0,95$ alors $1-\dfrac{\alpha}{2} = 0,975$ et $t_\alpha \simeq 1,96$.
- Comme mentionné précédemment, le niveau de confiance $1-\alpha = 0,95$ est assez classique. Avec un tel niveau de confiance, on considère qu'il y a 95% de chances de tomber sur un panel standard. Lorsque c'est le cas, le paramètre réel se retrouve dans l'intervalle $[\overline{x_n} - \varepsilon, \overline{x_n} + \varepsilon]$.
  - × On peut souhaiter obtenir un résultat très précis. Pour un sondage concernant des élections, savoir qu'un candidat est évalué à 19% plus ou moins 0,5% serait idéal. Du point de vue du sondeur, cela voudrait dire interroger $n = 38416$ personnes. C'est presque 6 fois moins que pour l'intervalle de confiance obtenu par inégalité de Bienaymé-Tchebychev. Pour autant, c'est toujours inenvisageable pour des raisons de coût.
  - × L'institut de sondage doit alors revoir ses objectifs à la baisse. En interrogeant $n = 1537$ personnes, il assure avec une probabilité de 95% qu'un candidat est évalué à 19% plus ou moins 2,5%. C'est 5 fois moins que dans le cas de l'intervalle de confiance obtenu par inégalité de Bienaymé-Tchebychev. Le coût est tout à fait envisageable et le résultat offre une précision correcte.
- Notons que les intervalles de confiance obtenus par les deux méthodes ont été réalisés avec la majoration : $p(1-p) \leqslant \dfrac{1}{4}$ $(*)$.
  - × La valeur $\frac{1}{4}$ est atteinte dans le cas où $p = \frac{1}{2}$. La majoration $(*)$ est donc la meilleure que l'on puisse faire en l'absence d'information sur p.
  - × Le rôle d'un sondage est justement d'obtenir de l'information sur p (une valeur approchée). Si un candidat est évalué à 20% (resp. 80%) alors on a $p(1-p) \simeq 0,16$.
    Avec ce calcul et pour $1-\alpha = 0,95$ et $n = 1500$, on obtient alors :

$$\varepsilon \;=\; \frac{\sqrt{p(1-p)}}{\sqrt{n}}\, t_\alpha \;\simeq\; \frac{\sqrt{0,16}}{\sqrt{1500}}\, 1,96 \;\simeq\; 0,02$$

    Ainsi, les sondages font souvent valoir une marge d'erreur qui dépend de l'estimation du candidat.