



17

STATISTIQUES

ESTIMATIONS PONCTUELLE ET PAR INTERVALLE DE CONFIANCE

INTRODUCTION...

Dans le chapitre 10, nous avons fourni des outils d'analyse de données numériques : moyenne, écart-type, médiane, quartiles, tableaux, graphiques... On se plaçait alors avec un certain nombre de données, que l'on analysait dans le but d'en extraire les informations essentielles : on parle de **statistique descriptive**.

Dans le chapitre qui suit, nous allons tenter, à partir de données obtenues sur un échantillon, d'en déduire des informations sur la population globale. On parle alors de **statistique inférentielle**. La statistique inférentielle est la statistique des sondages : on interroge un échantillon réduit de la population, et on tente d'estimer alors la tendance générale sur population entière. L'enjeu ici est de préciser la qualité des estimations proposées.

POUR BIEN DÉMARRER...

1. Rappeler l'inégalité de Bienaymé-Tchebychev.

Si X admet une variance, alors :

$$\forall a \in \mathbb{R}^{++}, \quad \mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathbb{V}(X)}{a^2}$$

2. Rappeler la loi faible des grands nombres.

Soit $(X_k)_{k \in \mathbb{N}^*}$ une suite de variables aléatoires sur $(\Omega, \mathcal{A}, \mathbb{P})$. Pour tout $n \in \mathbb{N}^*$, on note $\overline{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$.

Si $(X_k)_{k \in \mathbb{N}^*}$ est une suite de variables aléatoires **indépendantes**, admettant toutes la même espérance m et la même variance σ^2 (c'est le cas si elles ont toutes la même loi) alors :

$$\forall \varepsilon > 0, \quad \lim_{n \rightarrow +\infty} \mathbb{P}(|\overline{X}_n - m| \geq \varepsilon) = 0$$

3. Rappeler le théorème central limite.

Soit $(X_k)_{k \in \mathbb{N}^*}$ une suite de variables aléatoires **indépendantes, de même loi** (espérance m , variance σ^2 non nulle).

On pose, pour tout $n \in \mathbb{N}^*$:

$$\overline{X}_n = \frac{1}{n} \sum_{k=1}^n X_k \quad ; \quad \overline{X}_n^* = \frac{\overline{X}_n - \mathbb{E}(\overline{X}_n)}{\sigma(\overline{X}_n)} = \sqrt{n} \frac{\overline{X}_n - m}{\sigma}$$

Dans ce cas, la suite $(\overline{X}_n^*)_{n \in \mathbb{N}^*}$ converge en loi vers une variable aléatoire suivant la loi $\mathcal{N}(0; 1)$.

Autrement dit :

$$\forall x \in \mathbb{R}, \quad \lim_{n \rightarrow +\infty} \mathbb{P}(\overline{X}_n^* \leq x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

Ou encore :

$$\forall (a, b) \in \mathbb{R}^2, \quad \left(a < b \implies \lim_{n \rightarrow +\infty} \mathbb{P}(a \leq \overline{X}_n^* \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \right)$$

Lorsque n est *suffisamment grand*, on pourra considérer que : $\mathbb{P}(a \leq \overline{X}_n^* \leq b) \simeq \mathbb{P}(a \leq Z \leq b)$ où $Z \hookrightarrow \mathcal{N}(0; 1)$.

I INTRODUCTION

Plaçons nous dans la peau d'enquêteurs souhaitant réaliser un sondage sur l'avis des français concernant la qualité de l'enseignement privé par rapport à l'enseignement public.

On interroge alors des personnes, choisies de façon la plus aléatoire possible et en veillant à ce que cet échantillon soit le plus représentatif possible de la population française. À la question "Pensez-vous que l'on suit une meilleure scolarité dans un établissement privé que dans un établissement public ?", la personne sondée répond "oui" ou "non".

On attribue à "oui" la valeur 1 et à "non" la valeur 0.

Remarque

Le sondage permet de n'enquêter que sur un groupe restreint d'individus, par manque de temps ou de moyen pour interroger la population entière. Sinon, on organise un référendum.

On note θ la proportion de français votant "oui" et l'on considère donc une variable aléatoire X sur un certain espace probabilisable, telle que $X \hookrightarrow \mathcal{B}(\theta)$. **L'objectif est alors d'estimer au mieux la valeur de θ .**

Remarque

Majuscules pour les variables aléatoires, minuscules pour les réalisations.

Pour cela, revenons à notre sondage. On considère une suite de variables aléatoires $(X_k)_{k \in \mathbb{N}^*}$, toutes de même loi que X , telles que : pour tout $k \in \mathbb{N}^*$, X_k prend la valeur 1 si la k -ième personne interrogée répond "oui" à la question, et 0 sinon. Pour tout $k \in \mathbb{N}^*$, on notera x_k la réalisation de la variable aléatoire X_k dans le sondage effectué.

Le caractère aléatoire ne repose pas dans la réponse de l'individu choisi, mais bien dans le choix de l'individu (autrement dit, qu'une certaine personne fasse partie ou non des sondés). Bien évidemment, on comprend le risque de la méthode : choisir un échantillon qui ne soit pas suffisamment représentatif de la population et qui ne nous fournit donc pas une bonne estimation de l'avis général de la population.

L'objectif est de pouvoir conclure un sondage par une phrase du type :

"la valeur observée sur l'échantillon est *proche* de la valeur théorique de θ avec *tel niveau de confiance*"

L'enjeu mathématique est alors de préciser le "*proche*" et le "*tel niveau de confiance*" de la phrase précédente.

On considère un phénomène aléatoire, modélisé par une variable aléatoire X définie sur un espace probabilisable $(\Omega, \mathcal{A})$, dont la loi appartient à une famille de lois bien précise, dépendant d'un ou deux paramètres réels que l'on cherche à déterminer.

EXEMPLES 1

E1 On modélise le phénomène par une variable aléatoire X suivant une loi de Bernoulli, comme dans l'introduction ci-dessus, mais on ne connaît pas le paramètre de cette loi, que l'on note θ et qui est à chercher dans l'ensemble $\Theta = [0; 1]$.

E2 On modélise la durée de vie d'une ampoule par une variable aléatoire X suivant une loi exponentielle, dont on ne connaît pas le paramètre, noté θ , qui est à chercher dans l'ensemble $\Theta = \mathbb{R}^{+*}$.

E3 On modélise le poids des nouveaux-nés par une variable aléatoire X suivant une loi normale, dont on ne connaît pas le couple des paramètres, noté $\theta = (\theta_1, \theta_2)$, qui est à chercher dans l'ensemble $\Theta = \mathbb{R} \times \mathbb{R}^{+*}$.

En fait...

En pratique, on commence par les outils de statistique descriptive pour analyser les données et essayer d'imaginer la famille de lois modélisant la situation : normale, exponentielle, Poisson, log-normale, Pareto... Après avoir fixé une famille de lois, on cherche le ou les paramètres de cette dernière par les méthodes de statistique inférentielle.

On pourra proposer une unique valeur possible de ce paramètre, on parle alors d'**estimation ponctuelle**, ou bien tout un intervalle de valeurs possibles, on parle alors d'**estimation par intervalle de confiance**.

Dans tout le chapitre, n désignera un entier naturel non nul. On considèrera également une variable aléatoire X définie sur un espace probabilisable $(\Omega, \mathcal{A})$ dont la loi dépend d'un paramètre θ réel prenant ses valeurs dans un ensemble Θ . On munit l'espace probabilisable d'une famille de probabilités $(\mathbb{P}_\theta)_{\theta \in \Theta}$.

II ESTIMATION PONCTUELLE

II.1 ÉCHANTILLON ET ESTIMATEUR

DÉFINITION 1

ÉCHANTILLON

Un **n -échantillon de X** est un n -uplet $(X_1, \dots, X_n)$ de variables aléatoires définies sur $(\Omega, \mathcal{A})$ indépendantes et de même loi que X pour toutes les $\mathbb{P}_\theta$.

DÉFINITIONS 2

ESTIMATEUR, ESTIMATION

Soit $(X_1, \dots, X_n)$ un n -échantillon de X .

D1 Un **estimateur de θ** est une variable aléatoire sur $(\Omega, \mathcal{A})$ fonction des variables aléatoires $X_1, \dots, X_n$ et dont l'expression ne fait pas mention de θ .

Autrement dit, une variable aléatoire T_n sur $(\Omega, \mathcal{A})$ est un estimateur de θ lorsqu'il existe une fonction φ définie sur $X_1(\Omega) \times \dots \times X_n(\Omega)$ et à valeurs dans $\mathbb{R}$ telle que $T_n = \varphi(X_1, \dots, X_n)$.

D2 Soit $\varphi(X_1, \dots, X_n)$ est un estimateur de θ . Soit $(x_1, \dots, x_n) \in X_1(\Omega) \times \dots \times X_n(\Omega)$. On dit que $\varphi(x_1, \dots, x_n)$ est une **estimation ponctuelle de θ** .

Autrement dit, une estimation ponctuelle de θ est une réalisation d'un estimateur de θ .

Important !

La loi d'un estimateur de θ dépend naturellement de θ (puisque l'estimateur est fonction des X_k dont la loi dépend de θ) ; **mais son expression ne doit pas faire intervenir θ .**

EXEMPLE 2

Soit X une variable aléatoire suivant la loi $\mathcal{B}(\theta)$, avec $\theta \in]0; 1[$. On considère $X_1, \dots, X_n$ des variables aléatoires indépendantes suivant toutes la même loi que X et on note : $T_n = \frac{1}{n} \sum_{k=1}^n X_k$.

Justifions que T_n est un estimateur de θ .

On sait que :

- ✓ $X_1, \dots, X_n$ sont indépendantes et suivent toutes la même loi $\mathcal{B}(\theta)$,
- ✓ T_n est fonction de $X_1, \dots, X_n$ dont l'expression ne fait pas apparaître θ .

Par conséquent : T_n est un estimateur de θ .

De même : $\sum_{k=1}^n X_k, \max(X_1, \dots, X_n), \min(X_1, \dots, X_n)$ sont des estimateurs de θ .

En revanche, $\sum_{k=1}^n X_k - \theta$ n'est pas un estimateur de θ , car son expression fait intervenir θ .

Remarque

Ce n'est pas ce qui manque des estimateurs, il y en a une infinité : la définition est très générale. Mais lesquels sont des *bons* estimateurs ?

II.2 MOYENNE ET VARIANCE EMPIRIQUES

Dans cette partie, nous allons donner et étudier deux estimateurs usuels et pertinents, l'un pour l'espérance et l'autre pour la variance. Ces estimateurs peuvent ainsi être utilisés pour déterminer les paramètres des lois $\mathcal{B}(p)$, $\mathcal{P}(\lambda)$, $\mathcal{U}([1; n])$, $\mathcal{E}(\lambda)$, $\mathcal{U}([0; b])$, $\mathcal{N}(\mu; \sigma^2)$... puisque chacune des lois fait intervenir espérance et/ou variance en paramètres !

DÉFINITIONS 3

MOYENNE ET VARIANCE EMPIRIQUES

Soient X une variable aléatoire admettant une espérance et $(X_1, \dots, X_n)$ un n -échantillon de X .

D1 La **moyenne empirique** de $X_1, \dots, X_n$ est l'estimateur $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$.

D2 La **variance empirique** de $X_1, \dots, X_n$ est l'estimateur $V_n = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2$.

Remarque

C'est l'estimateur le plus naturel et le plus usuel de l'espérance d'une variable aléatoire ! Et on sait en plus qu'il est *bon*...

EXEMPLES 3

E1 Dans le cas du sondage initial, X suivait une loi de Bernoulli de paramètre p inconnu. Puisque $\mathbb{E}(X) = p$, on peut obtenir une estimation du paramètre p de la loi de Bernoulli en considérant la réalisation de la moyenne empirique sur un échantillon suffisamment grand.

Pour estimer $\mathbb{V}(X) = p(1-p)$, on pourrait par exemple proposer $\bar{X}_n(1 - \bar{X}_n)$...

E2 On suppose maintenant que X suit une loi exponentielle de paramètre λ inconnu.

On peut utiliser la moyenne empirique d'un échantillon pour obtenir une estimation de $\mathbb{E}(X) = \frac{1}{\lambda}$, ce qui nous fournit ensuite une estimation de λ ...

E3 On suppose maintenant que X suit une loi uniforme sur $[0; b]$, avec b inconnu.

On peut utiliser la moyenne empirique d'un échantillon pour obtenir une estimation de $\mathbb{E}(X) = \frac{b}{2}$, ce qui nous fournit ensuite une estimation de b ...

Remarque

Tout ce qui précède est à adapter dans le cas où l'on cherche à estimer $g(\theta)$, et pas θ (où g est une fonction définie sur Θ). De façon générale, si $\mathbb{E}(X) = f(\theta)$, où f est bijective sur Θ , alors une estimation de θ sera fournie par une réalisation de $f^{-1}(\bar{X}_n)$.

E4 Démontrons que $V_n = \frac{1}{n} \sum_{k=1}^n X_k^2 - \bar{X}_n^2$ puis déduisons-en l'espérance de V_n .

$$\begin{aligned} V_n &= \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2 \\ &= \frac{1}{n} \sum_{k=1}^n (X_k^2 - 2\bar{X}_n X_k + \bar{X}_n^2) \\ &= \frac{1}{n} \sum_{k=1}^n X_k^2 - 2\bar{X}_n \frac{1}{n} \sum_{k=1}^n X_k + \bar{X}_n^2 \\ &= \frac{1}{n} \sum_{k=1}^n X_k^2 - \bar{X}_n^2 \end{aligned}$$

↙ $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$

Or :

- ✓ pour tout $k \in [1; n]$, X_k^2 admet une espérance car X admet une variance ;
- ✓ $\bar{X}_n^2$ admet une espérance (car $\bar{X}_n$ admet une variance comme combinaison linéaire de telles variables aléatoires).

Par conséquent, V_n est une combinaison linéaire de variables aléatoires admettant une espérance ; V_n admet donc

une espérance et :

$$\begin{aligned}
 \mathbb{E}(V_n) &= \mathbb{E} \left(\frac{1}{n} \sum_{k=1}^n X_k^2 - \overline{X_n}^2 \right) && \leftarrow \text{linéarité de l'espérance, toutes existent} \\
 &= \frac{1}{n} \sum_{k=1}^n \mathbb{E}(X_k^2) - \mathbb{E}(\overline{X_n}^2) && \leftarrow \text{les } X_k \text{ ont même loi que } X \\
 &= \frac{1}{n} \sum_{k=1}^n \mathbb{E}(X^2) - \mathbb{E}(\overline{X_n}^2) \\
 &= \mathbb{E}(X^2) - \mathbb{E}(\overline{X_n}^2) \\
 &= \mathbb{V}(X) + \mathbb{E}(X)^2 - \mathbb{V}(\overline{X_n}) - \mathbb{E}(\overline{X_n})^2 && \leftarrow \text{formule de Koenig-Huygens} \\
 &= \frac{n-1}{n} \mathbb{V}(X) && \leftarrow \mathbb{E}(\overline{X_n}) = \mathbb{E}(X) \text{ et } \mathbb{V}(\overline{X_n}) = \frac{1}{n} \mathbb{V}(X)
 \end{aligned}$$

Conclusion : la variance empirique est un estimateur de $\mathbb{V}(X)$ d'espérance $\frac{n-1}{n} \mathbb{V}(X)$.

Remarque

Je ne détaille pas ces deux calculs, très classiques.

Pour info...

C'est pour cela qu'on préfère souvent utiliser $V'_n = \frac{1}{n-1} \sum_{k=1}^n (X_k - \overline{X_n})^2$; dont l'espérance est égale à $\mathbb{V}(X)$. V'_n est appelé **variance empirique corrigée** et est souvent utilisée dans les logiciels informatiques et calculatrices.

II.3 ESTIMATEUR DU MAXIMUM DE VRAISEMBLANCE

On considère toujours une variable aléatoire suivant une certaine loi, dont un paramètre est inconnu, et un n -échantillon $(X_1, \dots, X_n)$ de X ayant fourni un n -uplet de réalisations $(x_1, \dots, x_n)$.

Une idée assez naturelle consiste à considérer que la valeur de θ qui a permis de générer les observations est celle qui avait la plus grande probabilité de les générer.

EXEMPLE 4

Commençons par un cas simple où $n = 1$. On considère une variable aléatoire X suivant la loi $\mathcal{B}(15, p)$, avec p inconnu. Une réalisation de cette variable aléatoire fournit la valeur 5.

Question : quelle valeur de p maximise $\mathbb{P}([X = 5])$?

Puisque $X \hookrightarrow \mathcal{B}(15, p)$, on a :

$$\mathbb{P}([X = 5]) = \binom{15}{5} p^5 (1-p)^{10}$$

La fonction $f : x \mapsto \binom{15}{5} x^5 (1-x)^{10}$ est polynomiale, donc dérivable sur $[0; 1]$ et, pour tout $x \in [0; 1]$:

$$\begin{aligned}
 f'(x) &= \binom{15}{5} (5x^4(1-x)^{10} - 10x^5(1-x)^9) \\
 &= \binom{15}{5} 5x^4(1-x)^9(1-x-2x) \\
 &= \binom{15}{5} 5x^4(1-x)^9(1-3x)
 \end{aligned}$$

D'où :

x	0	$\frac{1}{3}$	1
$f'(x)$	+	0	-
f	0	$\nearrow \dots \searrow$	0

Conclusion : la probabilité $\mathbb{P}([X = 5])$ est maximale lorsque $p = \frac{1}{3}$.
Puisque l'évènement $[X = 5]$ s'est réalisé, on peut penser que $p = \frac{1}{3}$.

Remarque

Cette définition sera rappelée dans le sujet si besoin.

DÉFINITIONS 4

FONCTION DE VRAISEMBLANCE

Soit X une variable aléatoire suivant une loi dépendant d'un certain paramètre θ inconnu et $(X_1, \dots, X_n)$ un n -échantillon de X . Soit $(x_1, \dots, x_n) \in X_1(\Omega) \times \dots \times X_n(\Omega)$ un n -uplet d'observations.

D1 Si X est discrète, on appelle **fonction de vraisemblance** pour $(x_1, \dots, x_n)$ la fonction :

$$L_n : \theta \mapsto \prod_{k=1}^n \mathbb{P}_\theta([X_k = x_k])$$

Remarque

Dans le cas discret, on retrouve (par indépendance de $X_1, \dots, X_n$) :

$$L_n(\theta) = \mathbb{P} \left(\bigcap_{k=1}^n [X_k = x_k] \right)$$

D2 Si X est à densité, de densité f_θ , on appelle **fonction de vraisemblance** pour $(x_1, \dots, x_n)$ la fonction :

$$L_n : \theta \mapsto \prod_{k=1}^n f_\theta(x_k)$$

Puisqu'il est plus commode de dériver une somme qu'un produit (surtout pour les fonctions de deux variables), on travaille souvent sur la **log-vraisemblance**, égale à $\ln \circ L_n$. Les fonctions vraisemblance et log-vraisemblance ont mêmes variations et mêmes extrema (par stricte croissance de $\ln$)...

Sympa !

Le cadre du cours était celui d'un unique paramètre réel θ . La fonction L_n est ainsi une fonction d'une variable. En revanche, si l'on cherche les deux paramètres d'une loi, L_n sera une fonction de deux variables... Vous voyez les liens ?!

Pour info...

L'estimateur du maximum de vraisemblance est toujours convergent, mais pas nécessairement sans biais.

Pénible !

On utilise souvent la même lettre pour le paramètre à déterminer et pour la variable de la fonction de vraisemblance...

DÉFINITIONS 5

MAXIMUM DE VRAISEMBLANCE

Soit X une variable aléatoire suivant une loi dépendant d'un certain paramètre θ inconnu et $(X_1, \dots, X_n)$ un n -échantillon de X . Soit $(x_1, \dots, x_n) \in X_1(\Omega) \times \dots \times X_n(\Omega)$ un n -uplet d'observations. On note L_n la fonction de vraisemblance associée et on suppose qu'elle admet un maximum sur Θ , atteint en un unique réel.

D1 L'estimation du maximum de vraisemblance de θ est le réel θ_n^* en lequel L_n admet son maximum.

D2 Si g est la fonction telle que $\theta_n^* = g(x_1, \dots, x_n)$, alors l'estimateur du maximum de vraisemblance de θ est la variable aléatoire $T_n = g(X_1, \dots, X_n)$.

EXEMPLE 5

Déterminons l'estimateur du maximum de vraisemblance du paramètre p d'une loi de Bernoulli. Pour cela, considérons une variable aléatoire X suivant la loi $\mathcal{B}(p)$, avec $p \in]0; 1[$ inconnu. Considérons également, pour tout $n \in \mathbb{N}^*$, $(X_1, \dots, X_n)$ un n -échantillon de X (avec $n \in \mathbb{N}^*$) et $(x_1, \dots, x_n)$ une réalisation de cet échantillon. On pose

$$L_n : p \mapsto \prod_{k=1}^n \mathbb{P}(X_k = x_k)$$

Démontrons que la fonction L_n admet un maximum sur $]0; 1[$ atteint en un unique réel puis donnons l'estimateur du maximum de vraisemblance de p .

Pour tout $p \in]0; 1[$:

$$\begin{aligned} L_n(p) &= \prod_{k=1}^n \mathbb{P}(X_k = x_k) \\ &= \prod_{k=1}^n p^{x_k} (1-p)^{1-x_k} \end{aligned} \quad \left. \begin{array}{l} \forall k \in \llbracket 1; n \rrbracket, X_k \hookrightarrow \mathcal{B}(1; p) \end{array} \right\}$$

D'où, pour tout $p \in]0; 1[$, $L_n(p) > 0$ et :

$$\begin{aligned} \varphi_n(p) &= \ln(L_n(p)) \\ &= \ln \left(\prod_{k=1}^n p^{x_k} (1-p)^{1-x_k} \right) \\ &= \sum_{k=1}^n \ln(p^{x_k} (1-p)^{1-x_k}) \\ &= \sum_{k=1}^n (x_k \ln(p) + (1-x_k) \ln(1-p)) \\ &= \left(\sum_{k=1}^n x_k \right) \ln(p) + \left(n - \sum_{k=1}^n x_k \right) \ln(1-p) \\ &= s_n \ln(p) + (n - s_n) \ln(1-p) \end{aligned} \quad \left. \begin{array}{l} \forall k \in \llbracket 1; n \rrbracket, p^{x_k} (1-p)^{1-x_k} > 0 \\ \forall k \in \llbracket 1; n \rrbracket, p > 0, 1-p > 0 \end{array} \right\}$$

La fonction φ_n est donc dérivable sur $]0; 1[$ et, pour tout $p \in]0; 1[$:

$$\begin{aligned} \varphi'_n(p) &= s_n \frac{1}{p} - (n - s_n) \frac{1}{1-p} \\ &= \frac{s_n(1-p) - (n - s_n)p}{p(1-p)} \\ &= \frac{s_n - np}{p(1-p)} \end{aligned}$$

D'où :

p	0	$\frac{s_n}{n}$	1
$\varphi'_n(p)$	+	0	-
φ_n		$\nearrow \dots \searrow$	

La fonction φ_n admet donc un maximum, atteint uniquement en $\frac{s_n}{n} = \frac{1}{n} \sum_{k=1}^n x_k$.

Or, par stricte croissance de $\ln$ sur $\mathbb{R}^{+*}$, les fonctions φ_n et L_n ont les mêmes variations sur $]0; 1[$.

Par conséquent : la fonction L_n admet un maximum atteint uniquement en $\frac{1}{n} \sum_{k=1}^n x_k$.

Conclusion : l'estimateur du maximum de vraisemblance de p est $\frac{1}{n} \sum_{k=1}^n X_k = \overline{X}_n$.

Remarque

On retrouve la moyenne empirique... mais ce ne sera pas toujours le cas !

III ESTIMATION PAR INTERVALLE DE CONFIANCE

III.1 INTERVALLE DE CONFIANCE (EXACT)

DÉFINITION 6

INTERVALLE DE CONFIANCE (EXACT)

Soit X une variable aléatoire suivant une loi dépendant d'un certain paramètre θ inconnu. Soient $\alpha \in]0; 1[$ ainsi que $(U_n)_{n \in \mathbb{N}^*}$ et $(V_n)_{n \in \mathbb{N}^*}$ deux suites de variables aléatoires.

On dit que $[U_n, V_n]$ est un intervalle de confiance de θ au niveau de confiance $1 - \alpha$ (ou au risque α), lorsque :

- ✓ $(U_n)_{n \in \mathbb{N}^*}$ et $(V_n)_{n \in \mathbb{N}^*}$ sont deux suites d'estimateurs de θ telles que : $\forall n \in \mathbb{N}^*, \mathbb{P}_\theta([U_n \leq V_n]) = 1$;
- ✓ $\mathbb{P}_\theta([\theta \in [U_n, V_n]]) \geq 1 - \alpha$

✗ Attention !

θ n'est pas une variable aléatoire ! En revanche, $[U_n, V_n]$ est un intervalle aléatoire !
On garde en tête :

$[\theta \in [U_n, V_n]] = [U_n \leq \theta] \cap [V_n \geq \theta]$
et on voit mieux, sous cette forme, ce qui est aléatoire et ce qui ne l'est pas !

Si u_n et v_n sont des réalisations de U_n et V_n respectivement, il est faux de dire " θ est compris entre u_n et v_n avec une probabilité au moins égale à $1 - \alpha$ ".

En effet, rien d'aléatoire dans θ , u_n et v_n .

On dira : "on a une confiance de $1 - \alpha$ dans le fait que θ soit compris entre u_n et v_n ".

Voyons déjà un premier exemple :

EXEMPLE 6

On considère une variable aléatoire X suivant une loi $\mathcal{N}(\mu; \sigma^2)$, où m est inconnu et σ^2 connu non nul. On dispose d'un n -échantillon $(X_1, \dots, X_n)$ de X .

- Donnons la loi de $\overline{X}_n$ et celle de $\overline{X}_n^* = \sqrt{n} \frac{\overline{X}_n - m}{\sigma}$.

Puisque $X_1, \dots, X_n$ sont indépendantes et de même loi que X , on a :

$$\overline{X}_n \hookrightarrow \mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$$

Par conséquent :

$$\frac{\overline{X}_n - m}{\sqrt{\frac{\sigma^2}{n}}} \hookrightarrow \mathcal{N}(0; 1)$$

Autrement dit :

$$\overline{X}_n^* \hookrightarrow \mathcal{N}(0; 1)$$

- Soit $\alpha \in]0; 1[$. Justifions l'existence d'un unique réel t_α tel que $\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$, où Φ désigne la fonction de répartition d'une variable aléatoire suivant la loi $\mathcal{N}(0; 1)$. Montrons que $t_\alpha > 0$.

- * Notons φ la densité continue sur $\mathbb{R}$ associée à la loi $\mathcal{N}(0; 1)$. Puisque φ est continue sur $\mathbb{R}$, la fonction Φ est $\mathcal{C}^1$ sur $\mathbb{R}$ et pour tout $x \in \mathbb{R}$:

$$\Phi'(x) = \varphi(x) > 0$$

Par conséquent :

- ✓ Φ est continue sur $\mathbb{R}$,
- ✓ Φ est strictement croissante sur $\mathbb{R}$.

Donc par théorème de bijection, Φ est bijective de $\mathbb{R}$ dans $\Phi(\mathbb{R})$.

Or $\Phi(\mathbb{R}) =]0; 1[$.

Ensuite, puisque $\alpha \in]0; 1[$, on a $1 - \frac{\alpha}{2} \in \left] \frac{1}{2}; 1 \right[$.

Conclusion : il existe un unique réel t_α tel que $\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$.

- * On a vu que $1 - \frac{\alpha}{2} \in \left] \frac{1}{2}; 1 \right[$.

D'où :

$$\Phi(t_\alpha) > \Phi(0)$$

Et par stricte croissance de Φ sur $\mathbb{R}$, on obtient :

$$t_\alpha > 0$$

☞ Rappel...

Si $X_1, \dots, X_n$ sont indépendantes telles que pour tout $k \in \llbracket 1; n \rrbracket$, $X_k \hookrightarrow \mathcal{N}(m_k, \sigma_k^2)$,

alors $\sum_{k=1}^n a_k X_k$ suit la loi

$$\mathcal{N}\left(\sum_{k=1}^n a_k m_k; \sum_{k=1}^n a_k^2 \sigma_k^2\right).$$

Pourquoi ?

Car Φ est continue, strictement croissante, de limite 0 en $-\infty$ et 1 en $+\infty$.

♣ Méthode !

Dans ce cas de figure, pour comparer les antécédents, on compare les images...

- Dédoublons-en que $\left[\bar{X}_n - t_\alpha \frac{\sigma}{\sqrt{n}}, \bar{X}_n + t_\alpha \frac{\sigma}{\sqrt{n}} \right]$ est un intervalle de confiance de m au niveau de confiance $1 - \alpha$.

✓ En notant $U_n = \bar{X}_n - t_\alpha \frac{\sigma}{\sqrt{n}}$ et $V_n = \bar{X}_n + t_\alpha \frac{\sigma}{\sqrt{n}}$, on sait déjà que U_n et V_n sont des estimateurs de m (car fonctions d'un n -échantillon de X dont les expressions ne font pas apparaître m) et que $\mathbb{P}([U_n \leq V_n]) = 1$ (car $t_\alpha \frac{\sigma}{\sqrt{n}} > 0$).

✓ Montrons ensuite que $\mathbb{P} \left(\left[m \in \left[\bar{X}_n - t_\alpha \frac{\sigma}{\sqrt{n}}, \bar{X}_n + t_\alpha \frac{\sigma}{\sqrt{n}} \right] \right] \right) \geq 1 - \alpha$.

On a :

$$\begin{aligned}
 \mathbb{P} \left(\left[m \in \left[\bar{X}_n - t_\alpha \frac{\sigma}{\sqrt{n}}, \bar{X}_n + t_\alpha \frac{\sigma}{\sqrt{n}} \right] \right] \right) &= \mathbb{P} \left(\left[\bar{X}_n - t_\alpha \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X}_n + t_\alpha \frac{\sigma}{\sqrt{n}} \right] \right) \\
 &= \mathbb{P} \left(\left[-t_\alpha \frac{\sigma}{\sqrt{n}} \leq m - \bar{X}_n \leq t_\alpha \frac{\sigma}{\sqrt{n}} \right] \right) \quad \leftarrow \frac{\sigma}{\sqrt{n}} > 0 \\
 &= \mathbb{P} \left(\left[-t_\alpha \leq \frac{m - \bar{X}_n}{\frac{\sigma}{\sqrt{n}}} \leq t_\alpha \right] \right) \\
 &= \mathbb{P} \left(\left[t_\alpha \geq \sqrt{n} \frac{\bar{X}_n - m}{\sigma} \geq -t_\alpha \right] \right) \\
 &= \mathbb{P}([-t_\alpha \leq \bar{X}_n^* \leq t_\alpha]) \\
 &= \Phi(t_\alpha) - \Phi(-t_\alpha) \quad \leftarrow \bar{X}_n^* \hookrightarrow \mathcal{N}(0; 1), \text{ donc } \bar{X}_n^* \text{ est à densité} \\
 &= 2\Phi(t_\alpha) - 1 \quad \leftarrow \forall x \in \mathbb{R}, \Phi(-x) = 1 - \Phi(x) \\
 &= 1 - \alpha \quad \leftarrow \Phi(t_\alpha) = 1 - \frac{\alpha}{2}
 \end{aligned}$$

Conclusion : $\left[\bar{X}_n - t_\alpha \frac{\sigma}{\sqrt{n}}, \bar{X}_n + t_\alpha \frac{\sigma}{\sqrt{n}} \right]$ est un intervalle de confiance de m au niveau de confiance $1 - \alpha$.

Plus l'intervalle de confiance est réduit, plus le risque augmente; alors que plus l'intervalle de confiance a une grande amplitude, plus le risque est faible. En pratique, il faudra trouver un compromis entre les deux.

✉ Pour info...

En pratique, un risque de 0,05 est acceptable... on cherchera donc souvent un intervalle de confiance de niveau de confiance 0,95 (95%).

♣ MÉTHODE 1 ♣ Recherche d'un intervalle de confiance exact pour $\mathbb{E}(X)$ avec l'inégalité de Bienaymé-Tchebychev.

X admet une espérance m inconnue et une variance σ^2 . On considère T_n un estimateur de m tel que $\mathbb{E}(T_n) = m$ (estimateur sans biais). Résumons les étapes de l'obtention de l'intervalle de confiance :

1. Inégalité de Bienaymé-Tchebychev :

$$\forall \varepsilon > 0, \mathbb{P}(|T_n - m| \geq \varepsilon) \leq \frac{\mathbb{V}(T_n)}{\varepsilon^2}$$

2. Si $\mathbb{V}(T_n)$ dépend de m , on majore $\mathbb{V}(T_n)$ par v_n , indépendant de m (sinon, on prend $v_n = \mathbb{V}(T_n)$) et ainsi :

$$\forall \varepsilon > 0, \mathbb{P}(|T_n - m| \geq \varepsilon) \leq \frac{v_n}{\varepsilon^2}$$

3. Par passage à l'évènement contraire, on obtient :

$$\forall \varepsilon > 0, \mathbb{P}(|T_n - m| < \varepsilon) \geq 1 - \frac{v_n}{\varepsilon^2}$$

Autrement dit :

$$\forall \varepsilon > 0, \mathbb{P}([T_n - \varepsilon < m < T_n + \varepsilon]) \geq 1 - \frac{v_n}{\varepsilon^2}$$

4. Et puisque $[T_n - \varepsilon < m < T_n + \varepsilon] \subset [T_n - \varepsilon \leq m \leq T_n + \varepsilon]$, on obtient :

$$\forall \varepsilon > 0, \mathbb{P}([m \in [T_n - \varepsilon; T_n + \varepsilon]]) \geq 1 - \frac{v_n}{\varepsilon^2}$$

Reste à choisir ε tel que, pour le α donné, on ait $\frac{v_n}{\varepsilon^2} = \alpha$... et ainsi :

$$\mathbb{P}([m \in [T_n - \varepsilon; T_n + \varepsilon]]) \geq 1 - \alpha$$

★ Classique ! ★

C'est très classique (et le seul cas qui semble être au programme pour cette partie du cours). On prendra souvent (toujours ?) la moyenne empirique comme estimateur de la moyenne.

✗ Attention !

Le niveau de confiance ne peut pas dépendre du paramètre m que l'on cherche à estimer.

Remarque

L'intervalle de confiance proposé est centré en T_n , d'amplitude 2ε .

EXEMPLE 7

On considère une variable aléatoire suivant une loi de Bernoulli de paramètre p inconnu ainsi qu'un n -échantillon $(X_1, \dots, X_n)$ de X . On note $\bar{X}_n$ la moyenne empirique de cet échantillon.

- Montrons : $\forall \varepsilon > 0, \mathbb{P}(|\overline{X}_n - p| \geq \varepsilon) \leq 1 - \frac{1}{4n\varepsilon^2}$.
Soit $\varepsilon > 0$.

* On sait que $\overline{X}_n$ admet une espérance et une variance et que :

$$\mathbb{E}(\overline{X}_n) = p ; \quad \mathbb{V}(\overline{X}_n) = \frac{p(1-p)}{n}$$

* D'après l'inégalité de Bienaymé-Tchebychev, licite car $\overline{X}_n$ admet une variance et que $\varepsilon > 0$:

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}(\overline{X}_n)| \geq \varepsilon) \leq \frac{\mathbb{V}(\overline{X}_n)}{\varepsilon^2}$$

Autrement dit :

$$\mathbb{P}(|\overline{X}_n - p| \geq \varepsilon) \leq \frac{p(1-p)}{n\varepsilon^2}$$

Mais, on sait que la fonction $x \mapsto x(1-x)$ est majorée par $\frac{1}{4}$; d'où :

$$\mathbb{P}(|\overline{X}_n - p| \geq \varepsilon) \leq \frac{1}{4n\varepsilon^2}$$

Or :

$$\begin{aligned} \mathbb{P}(|\overline{X}_n - p| \geq \varepsilon) &= 1 - \mathbb{P}(|\overline{X}_n - p| < \varepsilon) \\ &= 1 - \mathbb{P}(\overline{X}_n - \varepsilon < p < \overline{X}_n + \varepsilon) \end{aligned} \quad \leftarrow \varepsilon > 0$$

D'où :

$$\mathbb{P}(\overline{X}_n - \varepsilon < p < \overline{X}_n + \varepsilon) \geq 1 - \frac{1}{4n\varepsilon^2}$$

Enfin, on a $[\overline{X}_n - \varepsilon < p < \overline{X}_n + \varepsilon] \subset [\overline{X}_n - \varepsilon \leq p \leq \overline{X}_n + \varepsilon]$ et par croissance de $\mathbb{P}$, on obtient par transitivité :

$$\mathbb{P}(\overline{X}_n - \varepsilon \leq p \leq \overline{X}_n + \varepsilon) \geq 1 - \frac{1}{4n\varepsilon^2}$$

Remarques

- En se fixant un risque $\alpha = \frac{1}{4n\varepsilon^2}$, plus ε diminue et plus n doit être grand pour conserver le même risque.
- En prenant $\alpha = \frac{1}{4n\varepsilon^2}$, on voit aussi que, à n fixé, plus ε est petit et plus α est grand... et plus ε est grand, plus α est petit.

- Déduisons-en que $\left[\overline{X}_n - \sqrt{\frac{5}{n}}, \overline{X}_n + \sqrt{\frac{5}{n}} \right]$ est un intervalle de confiance de p au niveau de confiance 95%.

✓ En notant $U_n = \overline{X}_n - \sqrt{\frac{5}{n}}$ et $V_n = \overline{X}_n + \sqrt{\frac{5}{n}}$, on sait déjà que U_n et V_n sont des estimateurs de p (car fonctions d'un n -échantillon de X dont les expressions ne font pas apparaître p) et que $\mathbb{P}(U_n \leq V_n) = 1$.

✓ Montrons ensuite que $\mathbb{P}\left(p \in \left[\overline{X}_n - \sqrt{\frac{5}{n}}, \overline{X}_n + \sqrt{\frac{5}{n}}\right]\right) \geq 0,95$.

On a :

$$\begin{aligned} \mathbb{P}\left(p \in \left[\overline{X}_n - \sqrt{\frac{5}{n}}, \overline{X}_n + \sqrt{\frac{5}{n}}\right]\right) &= \mathbb{P}\left(\overline{X}_n - \sqrt{\frac{5}{n}} \leq p \leq \overline{X}_n + \sqrt{\frac{5}{n}}\right) \quad \leftarrow \text{point précédent avec } \varepsilon = \sqrt{\frac{5}{n}} \\ &\geq 1 - \frac{1}{20} = 0,95 \end{aligned}$$

Conclusion : $\left[\overline{X}_n - \sqrt{\frac{5}{n}}, \overline{X}_n + \sqrt{\frac{5}{n}} \right]$ est un intervalle de confiance de p au niveau de confiance 95%.

- Au second tour d'une élection présidentielle, les citoyens ont le choix entre deux candidats A et B . Un institut de sondage réalise un sondage auprès de 2000 personnes, dont 1150 affirment vouloir voter pour le candidat A . Peut-on affirmer, avec un risque d'erreur de 5% que le candidat A sera élu ?

Considérons X une variable aléatoire suivant la loi de Bernoulli de paramètre p , où p est la probabilité que le candidat A soit élu.

Pour tout $k \in \mathbb{N}^*$, X_k est la variable aléatoire suivant la loi de Bernoulli de succès "la k -ième personne interrogée vote le candidat A ".

En supposant le panel du sondage bien choisi, on peut considérer que $(X_1, \dots, X_{2000})$ est un échantillon de X .

Notons $\overline{x}_{2000}$ la réalisation de $\overline{X}_{2000}$ obtenue sur cet échantillon : $\overline{x}_{2000} = \frac{1150}{2000}$.

Dans ce cas, d'après ce qui précède, on peut estimer que p se situe dans l'intervalle $\left[\overline{x}_{2000} - \sqrt{\frac{5}{2000}}, \overline{x}_{2000} + \sqrt{\frac{5}{2000}} \right]$

avec un niveau de confiance 95%.

Mais :

$$\begin{aligned} \left[\overline{x}_{2000} - \sqrt{\frac{5}{2000}}, \overline{x}_{2000} + \sqrt{\frac{5}{2000}} \right] &= \left[\frac{1150}{2000} - \sqrt{\frac{1}{400}}, \frac{1150}{2000} + \sqrt{\frac{1}{400}} \right] \\ &= [0,575 - 0,05; 0,575 + 0,05] \end{aligned}$$

$$= [0, 525; 0, 625]$$

On peut donc affirmer, avec un niveau de confiance de 95%, que la proportion de votes pour le candidat A se situe dans l'intervalle $[0, 525; 0, 625]$.

Conclusion : on peut affirmer, avec un risque d'erreur de 5%, que la candidat A sera élu.

III.2 INTERVALLE DE CONFIANCE ASYMPTOTIQUE

DÉFINITION 7

INTERVALLE DE CONFIANCE ASYMPTOTIQUE

Soit X une variable aléatoire suivant une loi dépendant d'un certain paramètre θ inconnu. Soient $\alpha \in]0; 1[$ ainsi que $(U_n)_{n \in \mathbb{N}^*}$ et $(V_n)_{n \in \mathbb{N}^*}$ deux suites de variables aléatoires.

On dit que $[U_n, V_n]$ est un **intervalle de confiance asymptotique de θ au niveau de confiance $1 - \alpha$ (ou au risque α)**, lorsque :

- ✓ $(U_n)_{n \in \mathbb{N}^*}$ et $(V_n)_{n \in \mathbb{N}^*}$ sont deux suites d'estimateurs de θ telles que : $\forall n \in \mathbb{N}^*, \mathbb{P}_\theta([U_n \leq V_n]) = 1$;
- ✓ $\lim_{n \rightarrow +\infty} \mathbb{P}_\theta([U_n, V_n]) \geq 1 - \alpha$

À retenir...

De façon plus générale, comme nous le verrons en exercice dans certains cas, une convergence en loi peut nous fournir un intervalle de confiance asymptotique.

♣ MÉTHODE 2 ♣ Recherche d'un intervalle de confiance asymptotique pour $\mathbb{E}(X)$ avec le TCL.

X admet une espérance m inconnue et une variance σ^2 non nulle. On travaille avec l'estimateur $\bar{X}_n$ de m (estimateur sans biais). Résumons les étapes de l'obtention de l'intervalle de confiance asymptotique :

- En posant $\bar{X}_n^* = \sqrt{n} \frac{\bar{X}_n - m}{\sigma}$, on a, par le TCL :

$$\bar{X}_n^* \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} Z$$

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

- Ainsi :

$$\forall x \geq 0, \lim_{n \rightarrow +\infty} \mathbb{P}([-x \leq \bar{X}_n^* \leq x]) = \mathbb{P}([-x \leq Z \leq x]) = 2\Phi(x) - 1$$

- Or :

$$\forall x \geq 0, \mathbb{P}([-x \leq \bar{X}_n^* \leq x]) = \dots = \mathbb{P}\left(\left[\bar{X}_n - \frac{\sigma x}{\sqrt{n}} \leq m \leq \bar{X}_n + \frac{\sigma x}{\sqrt{n}}\right]\right)$$

D'où :

$$\lim_{n \rightarrow +\infty} \mathbb{P}\left(\left[\bar{X}_n - \frac{\sigma x}{\sqrt{n}} \leq m \leq \bar{X}_n + \frac{\sigma x}{\sqrt{n}}\right]\right) = 2\Phi(x) - 1$$

- Si σ dépend de m , on majore σ par s , indépendant de m (sinon, on prend $s = \sigma$) et ainsi :

$$\left[\bar{X}_n - \frac{\sigma x}{\sqrt{n}} \leq m \leq \bar{X}_n + \frac{\sigma x}{\sqrt{n}}\right] \subset \left[\bar{X}_n - \frac{s x}{\sqrt{n}} \leq m \leq \bar{X}_n + \frac{s x}{\sqrt{n}}\right]$$

D'où, par croissance de $\mathbb{P}$:

$$\lim_{n \rightarrow +\infty} \mathbb{P}\left(\left[\bar{X}_n - \frac{s x}{\sqrt{n}} \leq m \leq \bar{X}_n + \frac{s x}{\sqrt{n}}\right]\right) \geq 2\Phi(x) - 1$$

- Reste à choisir x tel que, pour le α donné, on ait $2\Phi(x) - 1 = 1 - \alpha$.

Mais :

$$\begin{aligned} 2\Phi(x) - 1 = 1 - \alpha &\iff \Phi(x) = 1 - \frac{\alpha}{2} \\ &\iff x = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \end{aligned} \quad \left\{ \begin{array}{l} \Phi \text{ est bijective de } \mathbb{R} \text{ dans }]0; 1[\text{ et} \\ 1 - \frac{\alpha}{2} \in]0; 1[\end{array} \right.$$

En notant $t_\alpha = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$, on obtient :

$$\lim_{n \rightarrow +\infty} \mathbb{P}\left(\left[m \in \left[\bar{X}_n - t_\alpha \frac{s}{\sqrt{n}}, \bar{X}_n + t_\alpha \frac{s}{\sqrt{n}}\right]\right]\right) \geq 2\Phi(t_\alpha) - 1 = 1 - \alpha$$

★ Classique ! ★

C'est très classique (et le seul cas qui semble être au programme pour cette partie du cours).

✗ Attention !

L'intervalle de confiance est donné par deux estimateurs, qui ne peuvent pas dépendre du paramètre m recherché ! Autrement dit, l'expression des bornes de l'intervalle de confiance ne doit pas faire apparaître m !

📖 Pour info...

Avec $\alpha = 0,05$, on a :
 $\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \simeq 1,96$.

Remarque

L'intervalle de confiance asymptotique proposé est centré en $\bar{X}_n$, d'amplitude $2t_\alpha \frac{s}{\sqrt{n}}$.

📖 Pour info...

De façon générale, σ est inconnu et on peut alors utiliser un estimateur de l'écart-type dans l'intervalle de confiance asymptotique...

EXEMPLE 8

On considère une variable aléatoire suivant une loi de Bernoulli de paramètre p inconnu ainsi qu'un n -échantillon $(X_1, \dots, X_n)$ de X . On note $\bar{X}_n$ la moyenne empirique de cet échantillon.

- Justifions que :

$$\forall x \geq 0, \lim_{n \rightarrow +\infty} \mathbb{P}\left(\left[-x \leq \sqrt{n} \frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \leq x\right]\right) = 2\Phi(x) - 1$$

où Φ est la fonction de répartition d'une variable aléatoire suivant la loi $\mathcal{N}(0; 1)$.

On sait que :

- ✓ $(X_k)_{k \in \mathbb{N}^*}$ est une suite de variables aléatoires indépendantes, de même loi, admettant la même espérance p et même variance $p(1-p)$ non nulle ;
- ✓ pour tout $n \in \mathbb{N}^*$, $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$ et

$$\begin{aligned}\bar{X}_n^* &= \frac{\bar{X}_n - \mathbb{E}(\bar{X}_n)}{\sigma(\bar{X}_n)} \\ &= \sqrt{n} \frac{\bar{X}_n - p}{\sigma} \\ &= \sqrt{n} \frac{\bar{X}_n - p}{\sqrt{p(1-p)}}\end{aligned}$$

Ainsi, d'après le théorème central limite :

$$\bar{X}_n^* \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} Z$$

où Z est une variable aléatoire suivant la loi $\mathcal{N}(0; 1)$.

Par conséquent, pour tout $x \in \mathbb{R}^+$

$$\begin{aligned}\lim_{n \rightarrow +\infty} \mathbb{P}([-x \leq \bar{X}_n^* \leq x]) &= \mathbb{P}([-x \leq Z \leq x]) \\ &= \Phi(x) - \Phi(-x) \\ &= 2\Phi(x) - 1\end{aligned}$$

$\hookrightarrow Z \hookrightarrow \mathcal{N}(0; 1)$, donc Z est à densité
 $\hookrightarrow \Phi(-x) = 1 - \Phi(x)$

Conclusion : $\forall x \geq 0, \lim_{n \rightarrow +\infty} \mathbb{P} \left(\left[-x \leq \sqrt{n} \frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \leq x \right] \right) = 2\Phi(x) - 1.$

- Soit $\alpha \in]0; 1[$. Démontrons l'existence d'un unique réel t_α tel que $\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$. On donne : $t_{0,05} \simeq 1,96$.

Voir exemple 6.

- Déduisons-en un intervalle de confiance asymptotique de p au niveau de confiance 95%.

- ✓ D'après ce qui précède avec $x = t_{0,05} \geq 0$, on a :

$$\lim_{n \rightarrow +\infty} \mathbb{P} \left(\left[-t_{0,05} \leq \sqrt{n} \frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \leq t_{0,05} \right] \right) = 2\Phi(t_{0,05}) - 1 = 1 - 0,05 = 0,95$$

Or :

$$\begin{aligned}\mathbb{P} \left(\left[-t_{0,05} \leq \sqrt{n} \frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \leq t_{0,05} \right] \right) &= \mathbb{P} \left(\left[\bar{X}_n - t_{0,05} \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq p \leq \bar{X}_n + t_{0,05} \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right] \right) \\ &\leq \mathbb{P} \left(\left[\bar{X}_n - t_{0,05} \frac{1}{2\sqrt{n}} \leq p \leq \bar{X}_n + t_{0,05} \frac{1}{2\sqrt{n}} \right] \right) \\ &\leq \mathbb{P} \left(\left[\bar{X}_n - \frac{1}{\sqrt{n}} \leq p \leq \bar{X}_n + \frac{1}{\sqrt{n}} \right] \right)\end{aligned}$$

$\hookrightarrow p(1-p) \leq \frac{1}{4}$, d'où l'inclusion
d'événements et par croissance de $\mathbb{P}...$
 $\hookrightarrow t_{0,05} \leq 2$ et même raisonnement

On obtient donc :

$$\lim_{n \rightarrow +\infty} \mathbb{P} \left(\left[\bar{X}_n - \frac{1}{\sqrt{n}} \leq p \leq \bar{X}_n + \frac{1}{\sqrt{n}} \right] \right) \geq 0,95$$

- ✓ Enfin, $\bar{X}_n - \frac{1}{\sqrt{n}}$ et $\bar{X}_n + \frac{1}{\sqrt{n}}$ sont des estimateurs de p (car fonctions d'un n -échantillon de X dont les expressions ne font pas apparaître p) et $\mathbb{P} \left(\left[\bar{X}_n - \frac{1}{\sqrt{n}} \leq \bar{X}_n + \frac{1}{\sqrt{n}} \right] \right) = 1.$

Conclusion : l'intervalle $\left[\bar{X}_n - \frac{1}{\sqrt{n}}; \bar{X}_n + \frac{1}{\sqrt{n}} \right]$ est un intervalle de confiance asymptotique de p au niveau de confiance 95%.

- Quelle taille d'échantillon devons-nous avoir pour que l'intervalle de confiance asymptotique donné ait une amplitude inférieure ou égale à 0,1 ?

L'amplitude de $\left[\bar{X}_n - \frac{1}{\sqrt{n}}; \bar{X}_n + \frac{1}{\sqrt{n}} \right]$ est égale à $\frac{2}{\sqrt{n}}$. Or, pour tout $n \in \mathbb{N}^*$:

$$\frac{2}{\sqrt{n}} \leq 0,1 \iff n \geq 400$$

Conclusion : il faudrait un échantillon d'au moins 400 individus pour que l'amplitude de cet intervalle soit inférieure ou égale à 0,1.

- Comparons cet intervalle avec celui obtenu dans l'exemple 7.

- * Puisque $\sqrt{5} > 1$, on a :

$$\left[\bar{X}_n - \frac{1}{\sqrt{n}}; \bar{X}_n + \frac{1}{\sqrt{n}} \right] \subset \left[\bar{X}_n - \sqrt{\frac{5}{n}}; \bar{X}_n + \sqrt{\frac{5}{n}} \right]$$

- * En comparaison, il fallait un échantillon de taille au moins 2000 pour obtenir une amplitude inférieure ou égale à 0,1 avec l'intervalle de confiance exact obtenu dans l'exemple 7.

Remarque

Quand $n = 100$, on obtient un intervalle de confiance de la forme $[\bar{X}_{100} - 0,1; \bar{X}_{100} + 0,1]$ ce qui est relativement grand pour estimer un paramètre $p \in]0; 1[$.