

Chapitre 13 - variables aléatoires : convergence, approximation et estimation

Objectifs d'apprentissage, à la fin de ce chapitre, je sais :

- utiliser les inégalités de Markov et de Bienaymé-Tchebychev ☐
- utiliser et interpréter la loi faible des grands nombres ☐
- démontrer la convergence en loi de variables aléatoires discrètes ou à densité ☐
- utiliser le théorème limite central ☐
- utiliser et justifier qu'une variable aléatoire est un estimateur d'un paramètre ☐
- manipuler les intervalles de confiance ☐

1 Inégalités et loi faible des grands nombres

1.1 Inégalités de Markov et de Bienaymé-Tchebychev

<u>Propriété - inégalité de Markov :</u> soit X une variable aléatoire réelle à valeurs positives admettant une espérance, alors $\forall a > 0, \quad P(X \geq a) \leq \frac{E(X)}{a}$	<u>Exemples :</u> • si $X \hookrightarrow \mathcal{G}\left(\frac{1}{3}\right)$ alors $P(X \geq 10) \leq \frac{3}{10}$ avec $a = 10$ • si $X \hookrightarrow \mathcal{E}(5)$, $P(X \geq 2) \leq \frac{1}{10}$
<u>Cas particuliers (classiques) :</u> • si X admet une espérance, $\forall a > 0, \quad P(X \geq a) \leq \frac{E(X)}{a}$ • si X admet un moment d'ordre 2 : $\forall a > 0, \quad P(X^2 \geq a) \leq \frac{E(X^2)}{a}$	<u>Remarque :</u> avec de telles transformations on obtient des variables aléatoires à valeur positives qui permettent d'appliquer l'inégalité de Markov.

Remarque : ces inégalités peuvent servir à obtenir des ordres de grandeur de certaines probabilités. Parfois leur utilisation n'apporte que peu d'information. C'est le cas par exemple de l'inégalité de Markov pour les variables à densité, on dispose généralement de la valeur exacte avec la fonction de répartition, $P(X \geq a) = 1 - F_X(a)$ avec l'exemple choisi, $P(X \geq 2) = 1 - (1 - e^{-5 \times 2}) \approx 4,5 \times 10^{-5}$

<u>Propriété - inégalité de Bienaymé-Tchebychev :</u> soit X une variable aléatoire admettant un moment d'ordre 2, alors : $\forall \varepsilon > 0, \quad P(X - E(X) \geq \varepsilon) \leq \frac{V(X)}{\varepsilon^2}$	<u>Exemple :</u> si $X \hookrightarrow \mathcal{B}(n, p)$ alors $P(X - E(X) \geq \varepsilon) \leq \frac{npq}{\varepsilon^2}$ on peut alors jouer avec la valeur de ε en fonction du résultat souhaité.
--	---

Remarque : cette inégalité donne une information sur la probabilité d'écart à la moyenne car

$$|X - E(X)| \geq \varepsilon \Leftrightarrow X \leq E(X) - \varepsilon \quad \text{ou} \quad E(X) + \varepsilon \leq X \Leftrightarrow X \notin [E(X) - \varepsilon, E(X) + \varepsilon]$$

Ce résultat est une »inégalité de concentration«. Grâce à ce type d'inégalités, on peut mesurer ou « contrôler » la probabilité qu'un phénomène aléatoire se disperse, i.e. s'écarte de la moyenne soit d'un comportement standard. Un exemple classique d'utilisation est de faire en sorte, avec le choix de ε , que $\frac{V(X)}{\varepsilon^2}$ vaille 0,05 ; il y a alors moins de 5% de chances que $|X - E(X)|$ soit supérieur au ε choisi, i.e. plus de 95% de chances que X soit dans $[E(X) - \varepsilon, E(X) + \varepsilon]$

1.2 loi faible des grand nombres

<u>Propriété - loi faible des grands nombres :</u> soit $(X_k)_{k \in \mathbb{N}}$ une suite de variables aléatoires réelles indépendantes admettant une même espérance μ et une même variance σ^2 on pose : $\overline{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$ pour $n \in \mathbb{N}^*$, appelée moyenne empirique de $X_1, \dots, X_n$ alors, pour tout $\varepsilon > 0$: $P\left(\overline{X}_n - \mu \geq \varepsilon\right) \xrightarrow{n \rightarrow +\infty} 0$	<u>Remarque :</u> cela signifie que, en statistique, la moyenne empirique $\overline{x}_n = \frac{1}{n} \sum_{k=1}^n x_k$ d'une suite de réalisations $x_1, x_2, \dots$ de variables indépendantes d'une même loi tend vers l'espérance μ quand la taille de l'échantillon n tend vers $+\infty$ On va alors pouvoir utiliser $\overline{x}_n$ comme un estimateur (i.e. une valeur approchée, cf. plus bas) de μ C'est ce résultat que l'on utilise avec Python, par exemple pour déterminer une valeur approchée de l'espérance d'une variable aléatoire
---	--

2 Convergence en loi, théorème limite central

2.1 Convergence en loi

<u>Définition :</u> soit $(X_k)_{k \in \mathbb{N}}$ une suite de variables aléatoires réelles. On dit que $(X_k)_{k \in \mathbb{N}}$ converge en loi vers la variable aléatoire X si, en tout point x où F_X est continue : $F_{X_k}(x) = P(X_k \leq x) \xrightarrow{k \rightarrow +\infty} P(X \leq x) = F_X(x)$ on notera dans ce cas : $X_k \xrightarrow[k \rightarrow +\infty]{\mathcal{L}} X$
--

<u>Propriété - convergence en loi de variables discrètes :</u> si $(X_k)_{k \in \mathbb{N}}$ est une suite de variables aléatoires discrètes et X une variable aléatoire discrète, toutes à valeurs dans $\mathbb{Z}$. Alors $X_k \xrightarrow[k \rightarrow +\infty]{\mathcal{L}} X$ si, et seulement si : $\forall x \in \mathbb{Z}, \quad P(X_k = x) \xrightarrow[k \rightarrow +\infty]{} P(X = x)$

Exemples et remarques :

- 1) Dans le cas discret, nous utiliserons systématiquement la propriété.
- 2) Dans le cas (rare pour nous) où la fonction de répartition n'est pas continue, cette définition peut donner des résultats suprenants.
Par exemple, si $X_n = \frac{1}{n}$ (loi certaine) pour $n \in \mathbb{N}^*$, alors $X_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} 0$
On remarquera que $P(X_n \leq 0) = 0$ ne converge pas vers $P(X \leq 0) = 1$ mais F_X n'étant pas continue en 0, cela ne contredit pas la définition donnée.
- 3) On rappelle qu'une variable à densité a une fonction de répartition continue sur $\mathbb{R}$.
- 4) Cas classique : chaines de Markov.
Dans une chaine de Markov si, pour tout $i \in \llbracket 1, r \rrbracket$, $P(X_n = i) \xrightarrow[n \rightarrow +\infty]{} p_i \in [0, 1]$, alors :
 - a) le vecteur $(p_1, \dots, p_r)$ définit un état stable, on peut considérer une variable V telle que $P(V = i) = p_i$ pour tout $i \in \llbracket 1, r \rrbracket$
 - b) X_n converge en loi vers V d'après la propriété (cas discret).

Nous avons donc déjà rencontré (sans le savoir) de nombreuses convergences en loi.

5) Cas classique : *convergence d'une suite de loi binomiale vers une loi de Poisson*

soit $\lambda > 0$ et $X_n \hookrightarrow \mathcal{B}\left(n, \frac{\lambda}{n}\right)$ pour $n \in \mathbb{N}^*$

alors $(X_n)_{n \in \mathbb{N}^*}$ converge en loi vers une variable aléatoire X suivant la loi de Poisson $\mathcal{P}(\lambda)$:

$$X_n \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} X \hookrightarrow \mathcal{P}(\lambda)$$

Idée pour la démonstration : soit $n \in \mathbb{N}^*$ et $k \in \mathbb{N}$ alors

$$P(X_n = k) = \binom{n}{k} p^k q^{n-k} = \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(\frac{n-\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!} \left(\frac{1}{k}\right) \frac{\prod_{i=0}^{k-1} (n-i)}{n^n} \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

on montre alors (ce n'est pas immédiat) que : $\frac{\prod_{i=0}^{k-1} (n-i)}{n^n} \rightarrow 1$ et $\left(1 - \frac{\lambda}{n}\right)^{n-k} \rightarrow e^{-\lambda}$

2.2 Théorème limite central

Ce théorème est parfois aussi appelé : théorème de la limite centrée ou théorème central - limite

Propriété - théorème limite central (TLC) :

soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires **non constantes, indépendantes, de même loi, admettant un moment d'ordre 2** (on note $\mu = E(X_n)$ et $\sigma^2 = V(X_n)$)

on pose : $\forall n \in \mathbb{N}^*, \quad \overline{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$ (moyenne empirique) et $\overline{X}_n^* = \sqrt{n} \left(\frac{\overline{X}_n - \mu}{\sigma} \right)$

alors la suite des variables centrées réduites $(\overline{X}_n^*)_{n \in \mathbb{N}^*}$ converge en loi vers une variable aléatoire suivant la loi normale centrée réduite :

$$\overline{X}_n^* \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} T \hookrightarrow \mathcal{N}(0, 1)$$

Corollaire : avec les hypothèses et notations du théorème, cela entraîne,

pour tous a, b tels que $-\infty \leq a < b \leq +\infty$:

$$P(a < \overline{X}_n^* < b) \xrightarrow[n \rightarrow +\infty]{} \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt$$

et quand a, b sont réels, on peut écrire :

$$P(a \leq \overline{X}_n^* \leq b) \xrightarrow[n \rightarrow +\infty]{} \Phi(b) - \Phi(a)$$

Remarques :

- en français, le théorème s'énonce : la suite des moyennes empiriques, centrées réduites, associée à une suite de variables indépendantes de même loi admettant un moment d'ordre 2 converge en loi vers une variable suivant la loi normale centrée réduite.
- $\overline{X}_n^* = \sqrt{n} \left(\frac{\overline{X}_n - \mu}{\sigma} \right)$ puisque $E(\overline{X}_n) = \frac{1}{n} \sum_{k=1}^n X_k = \frac{1}{n} \times n\mu$ et $V(\overline{X}_n) = \frac{1}{n^2} \times n\sigma^2$ car les variables sont indépendantes (pour V) et de même loi.
- si $S_n = \sum_{k=1}^n X_k$, alors $S_n^* = \overline{X}_n^*$ et donc il est équivalent de dire que S_n^* converge en loi vers une variable suivant la loi $\mathcal{N}(0, 1)$
- on rappelle que $X \hookrightarrow \mathcal{N}(0, 1) \Leftrightarrow \sigma X + \mu \hookrightarrow \mathcal{N}(\mu, \sigma^2)$

2.3 Des approximations classiques à l'aide du théorème limite central

Approximation d'une loi binomiale par une loi normale

Soit n un entier, $p \in]0, 1[$ et $Z_n \hookrightarrow \mathcal{B}(n, p)$

on peut considérer que $Z_n = \sum_{i=1}^n X_i$ où $(X_i)_{i \in \mathbb{N}^*}$ est une suite de variables indépendantes, de même loi de Bernoulli $\mathcal{B}(p)$

Nota bene : on peut d'ailleurs retrouver que $E(Z_n) = \sum_{i=1}^n E(X_i)$ par linéarité et donc $E(Z_n) = np$

ainsi que $V(Z_n) = \sum_{i=1}^n V(X_i)$ par indépendance et donc $V(Z_n) = npq$

alors $Z_n^* = \frac{Z_n - np}{\sqrt{npq}}$ est la variable centrée réduite associée à Z_n

et $Z_n^* = \frac{n \left(\frac{Z_n}{n} - p \right)}{\sqrt{n} \sqrt{pq}} = \sqrt{n} \times \frac{\overline{X_n} - p}{\sqrt{pq}}$ car $\frac{Z_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i = \overline{X_n}$

de plus $\forall i \in \llbracket 1, n \rrbracket$, $E(X_i) = p$ et $V(X_i) = pq$ donc $Z_n^* = \overline{X_n}^*$

de fait le théorème limite central s'applique : $Z_n^* = \overline{X_n}^* \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} T \hookrightarrow \mathcal{N}(0, 1)$, et la loi de Z_n^* peut être approchée par $\mathcal{N}(0, 1)$ quand n est suffisamment grand.

or $X \hookrightarrow \mathcal{N}(0, 1) \Leftrightarrow \sigma X + \mu \hookrightarrow \mathcal{N}(\mu, \sigma^2)$ et $Z_n = \sqrt{npq} Z_n^* + np$ donc on considère que

pour n suffisamment grand, la loi $\mathcal{B}(n, p)$ peut être approchée par $\mathcal{N}(np, npq)$

Remarque : si une loi normale peut correspondre, il faut, entre autres, qu'espérance et variance correspondent à celles de $\mathcal{B}(n, p)$, donc c'est forcément $\mathcal{N}(np, npq)$

Approximation d'une loi de Poisson par une loi normale

Soit n est un entier, $c > 0$, $Z_n \hookrightarrow \mathcal{P}(cn)$

on peut considérer que $Z_n = \sum_{i=1}^n X_i$ où $(X_i)_{i \in \mathbb{N}^*}$ est une suite de variables indépendantes, de même loi de Poisson $\mathcal{P}(c)$ (on le démontre par récurrence, cf. exercice 3)

alors de manière analogue, $Z_n^* = \frac{Z_n - cn}{\sqrt{cn}}$ est la variable centrée réduite associée à Z_n

et $Z_n^* = \frac{n \left(\frac{Z_n}{n} - c \right)}{\sqrt{n} \sqrt{c}} = \sqrt{n} \times \frac{\overline{X_n} - c}{\sqrt{c}}$ car $\frac{Z_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i = \overline{X_n}$

de plus $\forall i \in \llbracket 1, n \rrbracket$, $E(X_i) = c$ et $V(X_i) = c$ donc $Z_n^* = \overline{X_n}^*$

de fait le théorème limite central s'applique : $Z_n^* = \overline{X_n}^* \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} T \hookrightarrow \mathcal{N}(0, 1)$, et la loi de Z_n^* peut être approchée par $\mathcal{N}(0, 1)$ quand n est suffisamment grand.

or $X \hookrightarrow \mathcal{N}(0, 1) \Leftrightarrow \sigma X + \mu \hookrightarrow \mathcal{N}(\mu, \sigma^2)$ et $Z_n = \sqrt{cn} Z_n^* + cn$ donc on considère que pour n suffisamment grand, la loi $\mathcal{P}(cn)$ peut être approchée par $\mathcal{N}(cn, cn)$ et donc

On pourra donc approcher la loi $\mathcal{P}(\lambda)$ par la loi normale $\mathcal{N}(\lambda, \lambda)$ pour de grandes valeurs de λ

$\triangle$ Attention, dans les deux cas, on peut parler de convergence avec la variable centrée réduite Z_n^* grâce au théorème limite central, mais pas pour Z_n car la loi normale qui est l'approximation de Z_n dépend de n à chaque fois. D'où le terme d'**approximation**.

3 Estimation

L'estimation est une démarche de statistique inférentielle, c'est-à-dire qu'elle consiste à deviner ou prédire le fonctionnement de processus aléatoires (comme le paramètre d'une loi) à partir de résultats statistiques.

Par exemple, il peut s'agir de déterminer la probabilité de faire pile avec une pièce (pas forcément équilibrée) ; prévoir le résultat d'une élection...

Plusieurs questions peuvent se poser : quel est le sens des résultats statistiques ? comment la taille de l'échantillon influence les résultats ? quelle confiance peut-on accorder à une estimation ?

3.1 Echantillon, estimateur

<u>Définitions</u> : on appelle échantillon la donnée de n variables aléatoires $X_1, \dots, X_n$ indépendantes de même loi.	<u>Exemple</u> : cela peut être • n lancers d'une pièce • 1 000 personnes lors d'un sondage
<u>Définitions</u> : si la loi commune aux variables d'un échantillon dépend d'un paramètre θ , et si on souhaite estimer $h_\theta = h(\theta)$ (où h est une fonction), on appelle estimateur de $h(\theta)$ toute fonction de $X_1, \dots, X_n$ du type $T_n = \varphi(X_1, \dots, X_n)$ telle que les réalisations de $t_n = \varphi(x_1, \dots, x_n)$ de T_n sont calculables à partir des seules réalisations observées $(x_1, \dots, x_n)$ de l'échantillon $(X_1, \dots, X_n)$ <u>en particulier, le calcul d'une réalisation de T_n ne doit pas dépendre de θ</u>	<u>Interprétation de la définition</u> : un estimateur est donc une fonction des résultats observés. Et la définition d'un estimateur traduit une intention de fournir une estimation. Avec les notations ci-contre, que l'estimateur T_n a pour but d'approcher $h(\theta)$, c'est-à-dire qu'on choisit de faire l'approximation $t_n \approx h_\theta$, où t_n est une réalisation de T_n . Se pose alors la question de la pertinence de ce choix, que l'on justifiera avec certaines propriétés du chapitre.

Remarques :

- sans le savoir, nous avons déjà fait de l'estimation, avec Python notamment. Nous avons déjà fait ce « choix » de considérer le résultat d'une simulation comme une valeur approchée, par exemple de l'espérance d'une variable aléatoire.
- la formulation « ne pas dépendre de θ » peut paraître contradictoire avec le fait que le paramètre θ est sous-jacent à l'expérience. L'idée est qu'on ne fera pas apparaître θ dans la formule de l'estimateur, qui ne sera que fonction des résultats de l'expérience.

3.2 Estimation ponctuelle

Le terme « ponctuel » signifie que l'on cherche ici à déterminer une valeur pour le paramètre recherché.

Exemple : la moyenne empirique

Jeu de pile ou face : on cherche à estimer la probabilité p d'obtenir « pile ».

On effectue n lancers, on note X_k la variable aléatoire égale à 1 si, pour le lancer k , la pièce tombe sur « pile » et à 0 sinon, donc $X_k \hookrightarrow \mathcal{B}(p)$ et on suppose que les lancers sont indépendants.

On note $T_n = \frac{1}{n} \sum_{k=1}^n X_k$, la **moyenne empirique** de $X_1, \dots, X_n$

alors $E(T_n) = p$ et la loi faible des grands nombres donne : $\forall \varepsilon > 0, \quad P(|T_n - p| > \varepsilon) \xrightarrow[n \rightarrow +\infty]{} 0$

donc si on fixe $\varepsilon > 0$, la probabilité qu'une réalisation t_n de T_n soit dans l'intervalle $[p - \varepsilon, p + \varepsilon]$ se rapproche de 1 quand $n \rightarrow +\infty$ ce qui justifie le choix d'approcher p par t_n et plus n est grand, plus cette estimation sera, en pratique, précise.

Généralisation de l'utilisation de la moyenne empirique :

on suppose que $X_1, \dots, X_n$ sont n variables aléatoires indépendantes de même loi admettant un moment d'ordre 2 et on cherche à estimer $\theta = E(X_1)$

alors la loi faible des grands nombres justifie de prendre comme estimateur la moyenne empirique :

$$T_n = \frac{1}{n} \sum_{k=1}^n X_k$$

Estimation par maximum de vraisemblance

Il s'agit avec cette méthode de s'intéresser aux valeurs les plus probables (et non à l'espérance par exemple) et ainsi de déterminer le maximum d'une fonction. Comme nous l'avons vu pour plusieurs variables à densité, la fonction densité peut présenter un « pic » (maximum) et on comprend qu'un échantillon d'une telle variable prendra plus facilement des valeurs autour du pic.

Cette méthode consiste donc en la recherche d'un maximum à partir de réalisations d'un échantillon. Voir l'exercice 8 pour un exemple.

3.3 Estimation par intervalle de confiance

Contrairement à la section précédente, on ne cherchera pas une valeur précise pour le paramètre recherché, mais à montrer qu'il se situe dans un intervalle de valeur avec un certain niveau de confiance (i.e. une probabilité minimale, par exemple 0,95 ou 0,99).

Nous détaillerons ici la démarche pour **deux cas classiques**,

Utilisation de l'inégalité de Bienaymé-Tchebychev :

on rappelle que $|X - E(X)| < \varepsilon \Leftrightarrow X - \varepsilon < E(X) < X + \varepsilon$ (*)

et puisque $P(|X - E(X)| \geq \varepsilon) = 1 - P(|X - E(X)| < \varepsilon)$

l'inégalité nous donne $1 - P(|X - E(X)| < \varepsilon) \leq \frac{V(X)}{\varepsilon^2}$

et donc

$$1 - \frac{V(X)}{\varepsilon^2} \leq P(|X - E(X)| < \varepsilon)$$

et même pour les variables à densité

$$1 - \frac{V(X)}{\varepsilon^2} \leq P(|X - E(X)| \leq \varepsilon)$$

c'est exactement (en se référant à (*)) dire que $E(X)$ se trouve dans l'intervalle $[X - \varepsilon, X + \varepsilon]$

avec une probabilité au moins égale à $1 - \frac{V(X)}{\varepsilon^2}$

Utilisation de l'approximation par une loi normale :

$$\text{si } X^* \hookrightarrow \mathcal{N}(0, 1) \text{ alors pour } \alpha > 0, \quad P(-\alpha \leq X^* \leq \alpha) = \Phi(\alpha) - \Phi(-\alpha)$$

avec les tables de valeurs de la fonction Φ on peut alors déterminer la probabilité qu'une loi centrée réduite se trouve dans un certain intervalle de valeurs $[-\alpha, \alpha]$

On se ramènera à ce résultat grâce au théorème limite central qui permet de montrer la convergence en loi de la variable centrée réduite associée à une moyenne empirique et on parlera d'intervalle de confiance asymptotique, pour « n grand » : la probabilité que $\overline{X}_n^*$ soit dans $[-\alpha, \alpha]$ tend vers $\Phi(\alpha) - \Phi(-\alpha)$ et sera donc approchée par cette valeur.

On peut ensuite se ramener à la variable initiale (la moyenne empirique) par des opérations :

$$-\alpha \leq \overline{X}_n^* \leq \alpha \Leftrightarrow \dots \Leftrightarrow \dots \leq \overline{X}_n \leq \dots$$

Définitions et cas pratiques

<u>Définition</u> : si h_θ est un paramètre à estimer et U et V deux estimateurs de h_θ si $P(U \leq h_\theta \leq V) \geq 1 - \alpha$ on dit que $[U, V]$ est un intervalle de confiance de h_θ au niveau de confiance $1 - \alpha$ si $\lim_{n \rightarrow +\infty} P(U_n \leq h_\theta \leq V_n) = P(U \leq h_\theta \leq V)$ et que $P(U \leq h_\theta \leq V) \geq 1 - \alpha$ on dit que $[U, V]$ est un intervalle de confiance asymptotique de h_θ (toujours au niveau de confiance $1 - \alpha$)	<u>Remarques</u> : dans la pratique on cherchera souvent des probabilités supérieures à 0,95 (soit au moins 95% de chances) ou 0,99 et donc des valeurs de α : $\alpha = 0,05$ ou respectivement $\alpha = 0,01$ on parle de « risque » pour α (risque de ne pas être dans l'intervalle recherché). comme évoqué plus haut, l'application du théorème limite central est un cas classique menant sur un intervalle de confiance asymptotique.
--	--

Exemple 1 : encadrement par l'inégalité de Bienaymé-Tchebychev

On considère n lancers indépendants d'une pièce dont la probabilité d'obtenir PILE est p , supposée inconnue.

on note $X_k = 1$ si le lancer k donne pile, $X_k = 0$ sinon, de sorte que $X_k \hookrightarrow \mathcal{B}(p)$

on note $T_n = \frac{1}{n} \sum_{k=1}^n X_k$, alors T_n est un estimateur de p et $E(T_n) = p$ et $V(T_n) = \frac{p(1-p)}{n}$

L'inégalité de Bienaymé-Tchebychev donne :

$$P(|T_n - p| \geq \varepsilon) \leq \frac{p(1-p)}{n\varepsilon^2} \quad \text{donc} \quad 1 - \frac{p(1-p)}{n\varepsilon^2} \leq P(|T_n - p| < \varepsilon) \leq P(|T_n - p| < \varepsilon)$$

$$\text{et } |T_n - p| \leq \varepsilon \Leftrightarrow T_n - \varepsilon \leq p \leq T_n + \varepsilon, \text{ donc } P(U_n \leq p \leq V_n) \geq 1 - \alpha_n \quad \text{avec} \quad \begin{cases} U_n = T_n - \varepsilon \\ V_n = T_n + \varepsilon \\ \alpha_n = \frac{p(1-p)}{n\varepsilon^2} \end{cases}$$

Cette inégalité montre que p est dans l'intervalle $[U_n, V_n]$, de longueur 2ε , avec une probabilité supérieure à $1 - \alpha_n$. $[U_n, V_n]$ est l'**intervalle de confiance** de p au **niveau de confiance** de $1 - \alpha_n$

On voudrait (pour améliorer la précision d'une part et le niveau de confiance d'autre part) :

- ε le plus petit possible pour que toute valeur dans $[U_n, V_n]$ donne une approximation de p
- α_n le plus petit possible, pour que la probabilité que p soit dans l'intervalle $[U_n, V_n]$ soit la plus grande possible.

En général, on cherche un intervalle de confiance à un risque (α) inférieur à 5% ou 1%, sachant que :

- à n fixé, **plus le risque α_n est faible, plus l'intervalle est grand.**
- à ε fixé, **plus le risque α_n est faible, plus n est grand.**

Pour le cas de la loi de Bernoulli, on utilise souvent la majoration :

$$\forall p \in]0, 1[\quad p(1-p) \leq \frac{1}{4} \quad \text{de sorte que : } \alpha_n \leq \frac{1}{4n\varepsilon^2}$$

Application numérique :

si on souhaite un intervalle d'amplitude $\varepsilon = \frac{1}{100}$, et un risque inférieur à 5%

on cherche alors à réaliser $\frac{1}{4n\varepsilon^2} \leq \frac{5}{100}$ donc $4n\varepsilon^2 \geq 20$ et donc $n \geq \frac{20}{4} \times (10^2)^2$

donc on pourra prendre $n \geq 50\,000$

Exemple 2 : encadrement par application du théorème limite central

On cherche à estimer le pourcentage p de votes « oui » à un référendum avec un sondage ; pour cela, on considère un n -échantillon $(x_1, \dots, x_n)$ et on calcule la fréquence de « oui ».

On modélise la situation par n variables aléatoires de Bernoulli indépendantes $X_1, \dots, X_n$ de paramètre p , $X_k = 1$ signifiant la $k^{\text{ième}}$ personne interrogée répond « oui ».

On note $T_n = \frac{1}{n} \sum_{k=1}^n X_k$ et on suppose que T_n^* peut être approchée par la loi normale centrée réduite (N.B. : une condition acceptable est par exemple : $np(1-p) \geq 10$).

alors $E(T_n) = p$ et $V(T_n) = \frac{pq}{n}$ ($q = 1 - p$) et donc $Z_n^* = \sqrt{n} \frac{Z_n - p}{\sqrt{pq}}$

donc $[|Z_n - p| \leq \varepsilon] = \left[|Z_n^*| \leq \frac{\varepsilon \sqrt{n}}{\sqrt{pq}} \right]$

si on considère $Z_n^* \hookrightarrow \mathcal{N}(0, 1)$, et si on pose $\beta_n = \frac{\varepsilon \sqrt{n}}{\sqrt{pq}}$ alors :

$$P(|Z_n - p| \leq \varepsilon) \approx P\left(-\frac{\varepsilon \sqrt{n}}{\sqrt{pq}} \leq Z_n^* \leq \frac{\varepsilon \sqrt{n}}{\sqrt{pq}}\right) = \Phi(\beta_n) - \Phi(-\beta_n) = 2\Phi(\beta_n) - 1$$

car $\Phi(-x) = 1 - \Phi(x)$ et donc $P(Z_n - \varepsilon \leq p \leq Z_n + \varepsilon) \approx 2\Phi(\beta_n) - 1$

donc avec les notations de l'exemple 1 :

$U_n = Z_n - \varepsilon$, $V_n = Z_n + \varepsilon$, $1 - \alpha_n = 2\Phi(\beta_n) - 1$ (niveau de confiance) et $\alpha_n = 2(1 - \Phi(\beta_n))$ (niveau de risque).

Application numérique 1 : on souhaite un intervalle d'amplitude $\varepsilon = 1\%$, et un niveau de confiance supérieur à 95%

On doit prendre β_n tel que $\Phi(\beta_n) \geq 0,975$, soit $\frac{\varepsilon}{\sqrt{n}} \sqrt{p(1-p)} \geq \Phi^{-1}(0,975)$

la table des valeurs de Φ donne : $\Phi^{-1}(0,975) \approx 1,96$

donc $\sqrt{n} \geq \frac{1,96}{0,01} \times \sqrt{p(1-p)}$ soit $n \geq 10000 \times (1,96)^2 \times p(1-p)$ et on ne connaît pas p

mais on sait que $p(1-p) \leq \frac{1}{4}$, donc l'inégalité est toujours vraie dès que : $n \geq 10000 \times \frac{(1,96)^2}{4}$

donc à peu près $n \approx 10000$, soit 5 fois moins que l'information donnée par l'inégalité de Bienaymé-Tchebychev.

Application numérique 2 : on souhaite un intervalle de confiance au niveau de 95% avec un échantillon de 1000 personnes (effectif usuel pour des sondages en France). Quelle valeur peut-on prendre pour l'amplitude de l'intervalle de confiance ?

Ici, on doit avoir encore $\beta_n \geq 1,96$ avec $n = 1000$ et on cherche la valeur de ε (on la souhaite qu'elle soit la plus petite possible).

et on remarque qu'on a toujours (quelle que soit la valeur de p) $\beta_n \geq \varepsilon \sqrt{4000}$, donc il suffit d'avoir $\varepsilon \sqrt{4000} \geq 1,96$ pour obtenir le résultat, soit $\varepsilon \approx 3\%$

Pour visualiser quelques résultats avec Python

Voici un lien sur Capytale (code de l'activité : b9e3-10089543) pour réaliser quelques simulations avec Python abordées en classe et qui illustrent des résultats du chapitre.