Estimation

On explore dans ce chapitre les liens déjà entrevus précédemment entre probabilités et statistiques.

Supposons qu'on gère un parc d'ordinateurs. On veut prévoir les pannes de ces ordinateurs (pour gérer les commandes de nouveaux ordinateurs par exemple). Il est impossible d'effectuer un calcul exact pour cette prévision ; on va alors s'en remettre à un calcul probabiliste.
On note alors X la durée de vie d'un ordinateur, et on suppose que X est une variable aléatoire suivant une loi déterminée à l'avance (par exemple, une loi exponentielle). Il faut donc, à l'aide d'observations, déterminer le paramètre de cette loi.

Une chose assez naturelle est alors la suivante : on sait que si $X \hookrightarrow \mathscr{E}(\theta)$, alors $\mathbb{E}(X) = 1/\theta$. Il suffit alors d'observer les durées de vie d'un grand nombre d'ordinateurs ; la moyenne de ces observations (qu'on a déjà nommée *moyenne empirique*) donnera alors une approximation de l'espérance de la loi recherchée. Ainsi, si m est la moyenne empirique issue de l'observation, et θ le paramètre inconnu de notre loi, on pourra considérer l'approximation $\theta \simeq 1/m$.
À l'aide d'observations statistiques, on a inféré les paramètres d'un modèle théorique : on parle de *statistique inférentielle*.

Ceci nous ouvre par la suite l'accès à de nouvelles quantités : variance de la durée de vie d'un ordinateur ; durée de vie d'un système composé de 10 ordinateurs, qui tombe en panne dès qu'un des ordinateurs qui le composent tombe en panne (cette durée de vie s'obtient alors comme un min de variables aléatoires indépendantes), etc.

On peut ensuite se poser des questions de précision de notre approximation : la moyenne empirique donne-t-elle une bonne approximation de l'espérance ? comment cela dépend-il du nombre d'observations ? combien d'observations doit-on faire pour être sûr d'approximer à 10% près ? ou d'obtenir une approximation à 10% près de manière très probable ? etc.

Ces questions mettent en jeu la notion d'intervalle de confiance.

1 Définitions

Pour pouvoir faire des statistiques (et en inférer des calculs de probabilités), il faut avoir à disposition plusieurs observations d'un même phénomène. Nous modélisons cela par n variables aléatoires indépendantes, identiquement distribuées.

L'ensemble de ces observations constitue un **échantillon** :

Définition 1. *Soit X une variable aléatoire.*

- *On appelle n-échantillon de la loi de X tout n-uplet $(X_1, \dots, X_n)$, où les X_i sont des variables iid, de même loi que X.*
- *On appelle réalisation de l'échantillon $(X_1, \dots, X_n)$ toute valeur $(x_1, \dots, x_n) = (X_1(\omega), \dots, X_n(\omega))$ de ce n-uplet, avec $\omega \in \Omega$.*

Ainsi, un échantillon est un n-uplet de variables aléatoires ; et une réalisation de cet échantillon est un n-uplet de réels.

Exemple 1. Si $X \hookrightarrow \mathscr{B}(p)$, $(X_1, \dots, X_{10})$, où les X_i sont iid et suivent $\mathscr{B}(p)$, est un 10-échantillon de la loi de X ; et $(0,1,1,0,0,1,0,1,1,0)$ est une réalisation de cet échantillon.
Si l'expérience aléatoire est un lancer de pièce, où on note 0 pour Pile et 1 pour Face, l'échantillon est construit en faisant 10 lancers successifs ; la réalisation précédente correspond à la séquence de tirages : P F F P P F P F F P.

On suppose maintenant que X (et donc tous les X_i) suivant une loi dépendant d'un paramètre θ (dans le cas de l'exemple précédent on aurait $\theta = p$). On note Θ l'ensemble des valeurs possibles de θ : dans le cas d'une loi de Bernoulli on aura par exemple $\Theta = [0,1]$.

On cherche, à partir d'une réalisation du n-échantillon, à estimer la valeur de θ : on a besoin pour cela de considérer une fonction des X_i. On définit alors une nouvelle variable aléatoire, appelée *estimateur* :

Définition 2. *On suppose que X suit une loi dépendant d'un paramètre* θ. *Soit* $(X_1,\dots,X_n)$ *un n-échantillon de la loi de X.*
On appelle estimateur de θ *toute variable aléatoire de la forme* $\varphi(X_1,\dots,X_n)$, *où*φ *est une fonction de* $\mathbb{R}^n$ *dans* $\mathbb{R}$ *indépendante de* θ.
En pratique une réalisation $\varphi(x_1,\dots,x_n)$ *fournira une valeur approché de* θ, *où de* $g(\theta)$ *ou g est une certaine fonction (par exemple,* $g(\theta) = \dfrac{1}{\theta}$*).*
On pourra dire indifféremment que $\varphi(X_1,\dots,X_n)$ *est un estimateur de* θ, *ou de* $g(\theta)$.

Exemple 2.

- Dans le cas de la loi exponentielle $\mathscr{E}(\theta)$, la moyenne empirique est un estimateur de $\dfrac{1}{\theta}$ (car elle produit une estimation de l'espérance de la loi).
- Dans le cas de l'exemple 1, on peut estimer p en effectuant la moyenne des X_i sur plusieurs observations. $\overline{X_n} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ (c'est la moyenne empirique !) et donc un estimateur de p.
- Dans le cas d'une loi uniforme sur $[0,\theta]$ (avec $\theta > 0$), $\max(X_1,\dots,X_n)$ est un estimateur de θ, dont on pourrait attendre qu'il donne une valeur approchée de θ.
  La moyenne empirique est aussi un estimateur de θ, il renverra une valeur proche de $\dfrac{\theta}{2}$ (espérance de la loi) pour n assez grand.

2 Estimateurs, et leurs caractéristiques

2.1 Un point de notation

Le cadre de ce cours est la recherche d'une loi de probabilité dépendant d'un paramètre θ. Si X est la variable aléatoire dont on cherche la loi (notée $\mathscr{L}(\theta)$ en général), on voit que les quantités probabilistes associées à X (par exemple $\mathbb{P}(X \in [0,1])$, $\mathbb{E}(X)$, $V(X)$) dépendent maintenant de θ.

À chaque valeur de θ correspond en fait une nouvelle mesure de probabilité $\mathbb{P}_\theta$. Par exemple, si on considère $X \hookrightarrow \mathscr{U}(\llbracket 1,n \rrbracket)$ de paramètre inconnu n, alors :

- $\mathbb{P}_3(X \leqslant 2) = \dfrac{2}{3}$ (car on suppose $X \hookrightarrow \mathscr{U}\big(\llbracket 1,3 \rrbracket\big)$;
- $\mathbb{P}_{10}(X \leqslant 2) = \dfrac{1}{5}$ (car on suppose $X \hookrightarrow \mathscr{U}\big(\llbracket 1,10 \rrbracket\big)$).

Par suite, les quantités issues des calculs probabilistes dépendront aussi de θ : il faudrait noter $\mathbb{E}_\theta(X)$ l'espérance de X, $V_\theta(X)$ sa variance, ... tout cela est assez lourd et n'a jamais été vu sur une épreuve de concours. Nous ne le ferons donc pas.
Dans la suite, on notera :

- X une variable aléatoire suivant une loi dépendant d'un paramètre inconnu θ ;
- $(X_1,\dots,X_n)$ n variables aléatoires iid de même loi que X ;
- T_n un estimateur de $g(\theta)$ (g étant une fonction quelconque).

2.2 Quelques exemples d'estimateurs

2.2.1 La moyenne empirique

C'est l'estimateur que nous rencontrerons le plus fréquemment. On rappelle que la moyenne empirique d'un n-échantillon $(X_1,\dots,X_n)$ est définie par : $\overline{X_n} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

On a alors $\mathbb{E}\left(\overline{X_n}\right) = \mathbb{E}(X)$. $\overline{X_n}$ donne une estimation de l'espérance de la loi de X ; on a donc accès à une quantité dépendant du paramètre recherché.
Par exemple si on considère la loi de Poisson $\mathscr{P}(\lambda)$, la moyenne empirique de n tirages indépendants aura pour espérance λ : $\overline{X_n}$ est un estimateur de λ.

2.2.2 Construction d'estimateurs à partir de max ou de min

Soit $X \hookrightarrow \mathscr{U}([a,b])$. Si on veut estimer b (la plus grande valeur possible de X) il peut être intéressant d'effectuer plusieurs observations et de retenir la plus grande valeur.
L'estimateur retenu pour b est alors $T_n = \max(X_1,\dots,X_n)$.
On verra en exercice qu'on n'a en fait pas $\mathbb{E}(T_n) = b$, mais $\mathbb{E}(T_n) = a + \dfrac{n}{n+1}(b-a)$.
On a par contre $\mathbb{E}(T_n) \xrightarrow{n\to+\infty} b$: on dit que l'estimateur est *asymptotiquement sans biais* (HP).

2.2.3 L'estimateur du maximum de vraisemblance

Il s'agit d'une méthode générique pour trouver un estimateur d'un paramètre inconnu.

Soit X une variable suivant une loi de paramètre θ. On effectue n observations de X notées $x_1, x_2, \dots, x_n$; on recherche quelle valeur de θ pourrait rendre compte le mieux possible de ces observations.
Une idée est alors (pour une variable discrète) de considérer le produit

$$L(x_1,\dots,x_n) = \mathbb{P}(X = x_1) \times \mathbb{P}(X = x_2) \times \cdots \times \mathbb{P}(X = x_n)$$

(qu'on appelle *vraisemblance*), et de regarder la valeur du paramètre θ qui le rend maximal. Cette valeur s'exprime en fonction des x_i.

Regardons ceci sur une loi de Poisson. Si $X \hookrightarrow \mathscr{P}(\lambda)$, on sait : $\forall\, k \in \mathbb{N}$, $\mathbb{P}(X = k) = e^{-\lambda}\dfrac{\lambda^k}{k!}$.
Si $x_1,\dots,x_n$ sont les observations effectuées, la vraisemblance est donnée par

$$L(x_1,\dots,x_n;\lambda) = e^{-n\lambda}\frac{\lambda^{x_1+x_2+\cdots+x_n}}{x_1!x_2!\dots x_n!}$$

On considère maintenant L comme fonction de la seule variable λ, et on cherche en quelle valeur de λ elle atteint sa valeur maximale. La mise en œuvre de cette méthode suppose ici l'existence et l'unicité du max, ce qui n'a aucune raison d'avoir lieu dans un cas général.

Par croissance de la fonction ln, on voit qu'il suffit pour cela de maximiser la fonction $\lambda \mapsto L(x_1,\dots,x_n;\lambda)$ qui s'écrit

$$\lambda \mapsto -n\lambda + (x_1 + x_2 + \cdots + x_n)\ln(\lambda) - \ln(x_1!x_2!\dots x_n!)$$

dont la dérivée par rapport à λ vaut $-n + \dfrac{x_1 + x_2 + \cdots + x_n}{\lambda}$. Par étude du signe de la dérivée, on voit que la quantité ln(L) atteint son maximum en l'unique réel $\lambda = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$.
L'estimateur du maximum de vraisemblance sera donc $T_n = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$... et on retrouve la moyenne empirique !

Exercice 1. Montrer que l'estimateur du maximum de vraisemblance pour une loi de Bernoulli $\mathscr{B}(p)$ est également la moyenne empirique.
On pourra remarquer que si $X \hookrightarrow \mathscr{B}(p)$ on a : $\forall\, x \in \{0,1\}$, $\mathbb{P}(X = x) = (1-p)^{1-x}p^x$.

Dans le cas d'une variable à densité, on définit aussi la vraisemblance par $L(x_1,\dots,x_n) = f(x_1) \times f(x_2) \times \cdots \times f(x_n)$, où f est une densité de X. La même procédure de maximisation s'applique ensuite.

2.2.4 D'autres exemples

On voit qu'il n'y a pas unicité de l'estimateur : plusieurs estimateurs peuvent estimer la même quantité avec plus ou moins de précision. L'estimateur du maximum de vraisemblance n'est pas forcément le plus performant. Nous verrons d'autres exemples en exercice.

2.3 « Performance » d'un estimateur : convergence, et contrôle de sa vitesse par Bienaymé-Tchebychev

Jusqu'à présent, notre idée d'approximation se base surtout sur l'espérance, dont on sait qu'elle n'est pas toujours représentative des valeurs prises par une variable.

Si T_n est un estimateur de la quantité $g(\theta)$ on souhaiterait en fait s'assurer que T_n prend, avec une probabilité élevée, des valeurs très proches de $g(\theta)$. La bonne définition est celle d'estimateur convergent :

Définition 3. *On dit que l'estimateur* T_n *de* $g(\theta)$ *est convergent ssi :*

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

Remarque 1. On pense immédiatement à la loi faible des grands nombres... ce dernier résultat montre effectivement que

Proposition 1. *Si* $(X_n)_{n\in\mathbb{N}}$ *est une suite de variables iid, de même loi qu'une variable* X ***admettant une variance****, alors la moyenne empirique* $\overline{X_n}$ *est un estimateur convergent de* $\mathbb{E}(X)$.

On définit en fait usuellement deux notions qui mesurent la performance d'un estimateur :

- Le *biais* d'un estimateur T_n (vu comme estimateur de $g(\theta)$) est la quantité $b(T_n) = \mathbb{E}(T_n) - g(\theta)$;
- Le *risque quadratique* d'un estimateur T_n (vu comme estimateur de $g(\theta)$) est la quantité $r(T_n) = \mathbb{E}\left(\big(T_n - g(\theta)\big)^2\right)$

Ces notions ne sont plus au programme. Elles permettent en fait de discuter la convergence d'un estimateur : en effet, l'inégalité de Markov donne, dès que $V(T_n)$ existe :

$$\mathbb{P}\big(|T_n - g(\theta)| > \varepsilon\big) = \mathbb{P}\big(\big(T_n - g(\theta)\big)^2 > \varepsilon^2\big) \leqslant \frac{\mathbb{E}\left(\big(T_n - g(\theta)\big)^2\right)}{\varepsilon^2} = \frac{r(T_n)}{\varepsilon^2}$$

Ainsi, si $r(T_n) \to 0$, T_n est un estimateur convergent.
Il existe un moyen simple de calculer le risque d'un estimateur. En appliquant Konig-Huygens, on peut écrire :

$$\begin{aligned} r(T_n) = \mathbb{E}\left(\big(T_n - g(\theta)\big)^2\right) &= V(T_n - g(\theta)) + \mathbb{E}\big(T_n - g(\theta)\big)^2 \\ &= V(T_n) + \big(\mathbb{E}(T_n) - g(\theta)\big)^2 \\ r(T_n) &= V(T_n) + b(T_n)^2 \end{aligned}$$

où on a utilisé l'invariance de la variance par translation, et la linéarité de l'espérance.
Ce résultat porte le nom de *décomposition biais-variance*. Il n'est plus au programme et devra être redémontré.

Un cas particulier fréquent est celui d'un estimateur sans biais (ie, de biais égal à 0). Dans ce cas on a $\mathbb{E}(T_n) = g(\theta)$, et dans le cas d'existence de la variance de T_n, Bienaymé-Tchebychev donne :

$$\mathbb{P}\big(|T_n - g(\theta)| > \varepsilon\big) \leqslant \frac{V(T_n)}{\varepsilon^2}$$

Dans ce cas, on obtient :

Proposition 2. *Si* $\mathbb{E}(T_n) = g(\theta)$ *et si* $V(T_n) \to 0$*, alors* T_n *est un estimateur convergent de* $g(\theta)$.

Ces résultats ne sont pas exigibles, mais les méthodes sont classiques et, dans un énoncé abordant ces thématiques, vous aurez à effectuer ces raisonnements et calculs.

3 Quelques exemples

Dans tout ce qui suit $(X_n)_{n\in\mathbb{N}^*}$ est une suite de variables iid de même loi que X ; et on notera $\overline{X_n}$ la moyenne empirique des X_i.

Exercice 1. Soit $X \hookrightarrow \mathscr{B}(p)$. Montrer que $\overline{X_n}$ est un estimateur de p. Majorer $\mathbb{P}\left(|\overline{X_n} - p| > \varepsilon\right)$ en fonction de n, p et ε.

Exercice 2. Soit $X \hookrightarrow \mathscr{U}([0,\theta])$. Montrer que $2\overline{X_n}$ est un estimateur de θ, d'espérance égale à θ. Calculer la variance de cet estimateur et en déduire que $\mathbb{P}\left(\left|2\overline{X_n} - \theta\right| > \varepsilon\right)$ est de limite nulle pour $n \to +\infty$.

Exercice 3. Soit $X \hookrightarrow \mathscr{U}([0,\theta])$, et $U_n = \max(X_1,\dots,X_n)$.

1. Montrer que U_n est un estimateur de θ.

On rappelle qu'on a alors $F_{M_n} = F^n$ *où F est la fonction de répartition de* $\mathscr{U}([0,\theta])$.

2. Donner une densité de M_n ; calculer $\mathbb{E}(M_n)$ et montrer que $V(M_n) = \dfrac{n\theta^2}{(n+1)^2(n+2)}$
3. Déterminer un λ_n réel tel que $\mathbb{E}(\lambda_n U_n) = \theta$. On note $V_n = \lambda_n U_n$; calculer $V(V_n)$ et montrer que V_n est un estimateur convergent de θ.

Exercice 4. Soit $X \hookrightarrow \mathscr{U}([0,\theta])$. On va construire l'estimateur du maximum de vraisemblance de θ.

1. Rappeler l'expression de la densité usuelle, notée f, de la loi $\mathscr{U}([0,\theta])$.
2. La fonction de vraisemblance $L : (x_1,\dots,x_n,\theta)$ est définie pour tous $(x_1,\dots,x_n) \in (\mathbb{R}_+^*)^n$ par

$$L(x_1,...,x_n,\theta) = \prod_{i=1}^{n} f(x_i)$$

   Montrer qu'on a, une observation $(x_1,...,x_n)$ étant fixée :

$$L(x_1,...,x_n,\theta) = \begin{cases} 0 \text{ si } \theta < \max(x_1,\dots,x_n) \\ \theta^{-n} \text{ si } \theta \geqslant \max(x_1,\dots,x_n) \end{cases}$$

3. Les x_i étant fixés, donner la représentation graphique de la fonction $\theta \mapsto L(x_1,...,x_n,\theta)$ sur $\mathbb{R}_+^*$.
4. En déduire que l'estimateur du maximum de vraisemblance est donné par $\max(X_1,\dots,X_n)$.

Exercice 5. Soit $\theta > 0$, et X de densité $f(x) = \begin{cases} 0 \text{ si } x < \theta \\ e^{\theta - x} \text{ si } x \geqslant \theta \end{cases}$.

1. Montrer que f est bien une densité de probabilité.
2. Reconnaître la loi suivie par $Y = X - \theta$. Préciser $\mathbb{E}(Y)$ et $V(Y)$; en déduire $\mathbb{E}(X)$ et $V(X)$.
3. Soit $S_n = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - 1)$. Montrer que S_n est un estimateur de θ. Calculer $\mathbb{E}(S_n)$ et $V(S_n)$. Majorer $\mathbb{P}(|S_n - \theta| > \varepsilon)$ pour $\varepsilon > 0$ et en déduire que cette proba est de limite nulle.

Exercice 6. (Un estimateur convergent de la variance d'une loi normale).
Soit (X_i) un échantillon de la loi $\mathscr{N}(m,\sigma^2)$. On suppose m connu, et on cherche à estimer σ^2. On pose

$$S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$$

1. Montrer que S_n est un estimateur de σ^2. Donner la loi suivie par $X_i - m$, et en déduire que $\mathbb{E}(S_n) = \sigma^2$.
2. Montrer que $(X_i - m)^2$ admet une variance. En déduire que S_n admet une variance, et que $\lim\limits_{n\to+\infty} V(S_n) = 0$.
3. Montrer que S_n est un estimateur convergent de σ^2.

4 Estimation par intervalle

On s'est pour l'instant intéressés à des propriétés probabilistes des estimateurs (espérance et variance), et à des indicateurs de performance faisant intervenir une limite $n \to +\infty$ où n est le nombre d'observations.
En pratique, le nombre d'observations est limité ; et à n fixé on souhaite disposer de résultats permettant de quantifier la précision des estimations effectuées.

Une manière d'agir est de donner un *intervalle de confiance*, c'est-à-dire un intervalle, fonction de T_n, dans lequel la quantité à estimer se trouvera avec une probabilité donnée (et qu'on espère proche de 1). En pratique, on peut par exemple considérer un intervalle de confiance à 95% : l'intervalle $[T_n - \varepsilon_n, T_n + \varepsilon_n]$ sera un intervalle de confiance de $g(\theta)$ à 95% ssi $P\left(g(\theta) \in [T_n - \varepsilon_n, T_n + \varepsilon_n]\right) \geqslant 0,95$.

On définira aussi un *intervalle de confiance asymptotique* en faisant intervenir une limite $n \to +\infty$.
Typiquement, un intervalle de confiance s'obtiendra à l'aide de l'inégalité de Bienaymé-Tchebychev, alors qu'un intervalle de confiance asymptotique sera plutôt une conséquence du TCL.

4.1 Intervalle de confiance

On considère comme précédemment une v.a. X suivant la loi $\mathscr{L}(\theta)$, où θ est un paramètre à déterminer, à l'aide d'un estimateur T_n de $g(\theta)$ (g étant une fonction quelconque).

Un intervalle de confiance est un intervalle dont les bornes sont deux estimateurs (donc des variables aléatoires) ; les observations permettent de donner un tel intervalle, et de mesurer la probabilité que la quantité à estimer soit dans cet intervalle. Cette probabilité est appelée *niveau de confiance*.

Définition 4. *Soient* (U_n) *et* (V_n) *deux suites d'estimateurs de* $g(\theta)$*, tels que* $\mathbb{P}_\theta(U_n \leqslant V_n) = 1$*. On dit que* $[U_n, V_n]$ *est un intervalle de confiance de* $g(\theta)$ *au niveau de confiance* $1-\alpha$ *(avec* $\alpha \in [0,1]$*) ssi*

$$\forall\, \theta \in \Theta,\ \mathbb{P}\left(U_n \leqslant g(\theta) \leqslant V_n\right) \geqslant 1-\alpha$$

Remarque 2. Dans cette définition, α est voué à être petit (pour obtenir une confiance élevée). Par exemple, pour $\alpha = 0,05$, on obtient un intervalle de confiance à 95%.

4.2 Exemple : estimation d'une loi de Bernoulli

On reprend dans cette section l'estimation du paramètre p d'une loi $\mathscr{B}(p)$; l'estimateur est ici donné par la moyenne empirique $T_n = \overline{X_n}$.
Comme $\mathbb{E}(T_n) = p$, on essaie de voir dans quelle mesure une réalisation de T_n va donner une valeur proche de p. Pour estimer l'écart de T_n par rapport à p (qui est son espérance), on peut utiliser l'inégalité de Bienaymé-Tchebychev, qui donne :

$$\forall\, t > 0,\ \mathbb{P}\left(|T_n - p| > t\right) \leqslant \frac{V(T_n)}{t^2}$$

Ceci peut se réécrire

$$\mathbb{P}\left(|T_n - p| \leqslant t\right) \geqslant 1 - \frac{V(T_n)}{t^2}$$

soit encore

$$\mathbb{P}\left(p \in [T_n - t, T_n + t]\right) \geqslant 1 - \frac{p(1-p)}{nt^2}$$

Si on veut un intervalle de confiance $1-\alpha$, on cherche donc t tel que $\alpha = \dfrac{p(1-p)}{nt^2}$: ceci donne $t = \sqrt{\dfrac{p(1-p)}{n\alpha}}$.
On obtient :

$$\mathbb{P}\left(p \in \left[T_n - \sqrt{\frac{p(1-p)}{n\alpha}}, T_n + \sqrt{\frac{p(1-p)}{n\alpha}}\right]\right) \geqslant 1-\alpha$$

Ceci est toutefois peu satisfaisant car cet « intervalle de confiance » dépend de la valeur de p, qui est justement le paramètre que l'on cherche à déterminer. Il faut pouvoir trouver un intervalle valable pour tout p de $[0,1]$: autrement dit un ε_n tel que

$$\forall\, p \in [0,1],\ \left[T_n - \sqrt{\frac{p(1-p)}{n\alpha}}, T_n + \sqrt{\frac{p(1-p)}{n\alpha}}\right] \subset [T_n - \varepsilon_n, T_n + \varepsilon_n]$$

Une étude de fonction montre aisément que : $\forall\, p \in [0,1],\ 0 \leqslant p(1-p) \leqslant \dfrac{1}{4}$; de sorte que $\varepsilon_n = \sqrt{\dfrac{1}{4n\alpha}} = \dfrac{1}{2\sqrt{n\alpha}}$ convient.

On a finalement montré :

Si T_n est la moyenne empirique de $X \hookrightarrow \mathscr{B}(p)$, $\left[T_n - \dfrac{1}{2\sqrt{n\alpha}}, T_n + \dfrac{1}{2\sqrt{n\alpha}}\right]$ est un intervalle de confiance de niveau $1-\alpha$ pour p.

Si on cherche spécifiquement un intervalle à 95%, il faut prendre $\alpha = 0{,}05 = \dfrac{1}{20}$; on obtient alors l'intervalle $\left[T_n - \sqrt{\dfrac{5}{n}}, T_n + \sqrt{\dfrac{5}{n}}\right]$.

Remarque 3. Ici, les « deux estimateurs » sont en fait construits à partir d'un seul estimateur (ici T_n), autour duquel on se donne une précision donnée (ici $\dfrac{1}{2\sqrt{n\alpha}}$). Ce sera presque toujours le cas.

4.3 Intervalle de confiance asymptotique

Le TCL peut aussi intervenir dans ce genre de calculs : il justifie qu'à la limite $n \to +\infty$, on puisse utiliser la loi normale, ce qui fournit une précision bien supérieure à Bienaymé-Tchebychev. En contrepartie, on n'obtiendra pas une valeur exacte de la probabilité d'être hors de l'intervalle de confiance ; ceci ne se fera que dans la limite $n \to +\infty$. On parlera d'*intervalle de confiance asymptotique* : la définition est assez formelle.

Définition 5. *Soient* (U_n) *et* (V_n) *deux suites d'estimateurs de* $g(\theta)$*, tels que* $\mathbb{P}_\theta(U_n \leqslant V_n) = 1$*. On dit que* $[U_n, V_n]$ *est un intervalle de confiance asymptotique de* $g(\theta)$ *au niveau de confiance* $1-\alpha$ *ssi pour tout* $\theta \in \Theta$ *il existe une suite* $(\alpha_n) \to \alpha$ *telle que*

$$\forall\, n \in \mathbb{N},\ \mathbb{P}\left(U_n \leqslant g(\theta) \leqslant V_n\right) \geqslant 1 - \alpha_n$$

En pratique (c'est la forme de résultat que fournira le TCL) on privilégiera la définition suivante :

Définition 6. *Soient* (U_n) *et* (V_n) *deux suites d'estimateurs de* $g(\theta)$*, tels que* $\mathbb{P}_\theta(U_n \leqslant V_n) = 1$*. On dit que* $[U_n, V_n]$ *est un intervalle de confiance asymptotique de* $g(\theta)$ *au niveau de confiance* $1-\alpha$ *ssi*

$$\lim_{n \to +\infty} \mathbb{P}\left(U_n \leqslant g(\theta) \leqslant V_n\right) = 1 - \alpha$$

4.4 Intervalle de confiance asymptotique quand la variance est inconnue

Si on ne dispose pas de formule donnant la variance en fonction du paramètre à estimer, la méthode précédente est difficile à mettre en œuvre. On peut alors appliquer la procédure suivante.
On introduit la *variance empirique* obtenue à partir des moyennes empiriques :

$$\overline{S_n} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X_n}\right)^2$$

On peut alors l'utiliser pour former un intervalle de confiance pour l'espérance lorsque la variance est aussi inconnue.

Théorème 3. *Soit* X *une variable aléatoire d'espérance* μ *et de variance non nulle inconnue et* $(X_1,\dots,X_n)$ *un* n*-échantillon de* X*. Alors, l'intervalle*

$$\left[\overline{X_n} - t_\alpha \frac{\overline{S_n}}{\sqrt{n}}, \overline{X_n} + t_\alpha \frac{\overline{S_n}}{\sqrt{n}}\right]$$

où $\Phi(t_\alpha) = 1-\alpha/2$*, est un intervalle de confiance asymptotique pour* m *au niveau de confiance* $1-\alpha$*.*

5 Exercices

Exercice 7. Soit $X \hookrightarrow \mathscr{B}(p)$ et $(X_1,...,X_n)$ un n-échantillon de variables iid de même loi que X. Soit $\overline{X_n}$ la moyenne empirique des X_i, et X_n^* la moyenne empirique centrée réduite. On note $q = 1-p$.

1. Donner l'expression de $\overline{X_n}^*$ en fonction de $\overline{X_n}$, n, p et q.
2. En déduire que
$$\lim_{n\to+\infty} \mathbb{P}\left(|\overline{X_n} - p| \leqslant \varepsilon\sqrt{\frac{pq}{n}}\right) = 2\Phi(\varepsilon) - 1$$
3. Montrer que Φ est une bijection de $\mathbb{R}$ sur $]0,1[$.
4. Soit $\alpha \in]0,1[$. Montrer qu'il existe un unique $t_\alpha > 0$ tel que $\Phi(t_\alpha) = 1 - \dfrac{\alpha}{2}$.

   (on dit que où t_α *est le quantile d'ordre* $1-\dfrac{\alpha}{2}$ *de la loi* $\mathscr{N}(0,1)$ *; ce* t_α *apparaît très fréquemment dans ce genre de problème).*

   En déduire que
$$\lim_{n\to+\infty} \mathbb{P}\left(\overline{X_n} - t_\alpha\sqrt{\frac{pq}{n}} \leqslant p \leqslant \overline{X_n} + t_\alpha\sqrt{\frac{pq}{n}}\right) = 1-\alpha$$
5. On rappelle : $\forall\, p \in [0,1],\ p(1-p) \leqslant \dfrac{1}{4}$. Montrer l'inclusion entre événements :
$$\left(\overline{X_n} - t_\alpha\sqrt{\frac{pq}{n}} \leqslant p \leqslant \overline{X_n} + t_\alpha\sqrt{\frac{pq}{n}}\right) \subset \left(\overline{X_n} - t_\alpha\frac{1}{2\sqrt{n}} \leqslant p \leqslant \overline{X_n} + t_\alpha\frac{1}{2\sqrt{n}}\right)$$
   En déduire que $\left[\overline{X_n} - t_\alpha\dfrac{1}{2\sqrt{n}}, \overline{X_n} + t_\alpha\dfrac{1}{2\sqrt{n}}\right]$ est un intervalle de confiance asymptotique au niveau $1-\alpha$.
6. On considère $n = 1000$ et $\alpha = 0.05$. Quelle est la largeur de cet intervalle ? Comparer avec l'intervalle de confiance $\left[\overline{X_n} - \dfrac{1}{2\sqrt{n\alpha}}, \overline{X_n} + \dfrac{1}{2\sqrt{n\alpha}}\right]$ obtenu par Bienaymé-Tchebychev (voir cours).

Exercice 8. Soit $X \hookrightarrow \mathscr{N}(\theta,1)$; on cherche à estimer θ par la moyenne empirique.

1. Montrer que $\overline{X_n}$ suit une loi normale dont on donnera les paramètres.
2. En déduire que $\overline{X_n}^*$ suit une loi normale centrée réduite.
3. Soit $a > 0$. Calculer $\mathbb{P}\left(\overline{X_n} - \dfrac{a}{\sqrt{n}} \leqslant \theta \leqslant \overline{X_n} + \dfrac{a}{\sqrt{n}}\right)$. En déduire un intervalle de confiance de θ au niveau de confiance 0,95 (on donne : $\Phi^{-1}(0.975) \simeq 1.96$, cf. table de valeurs de Φ.)

Intervalle de confiance de $\mathscr{U}([0,\theta])$ avec Bienaymé-Tchebychev : cas de la moyenne empirique.
On pose $(X_1,\dots,X_n)$ un échantillon de $\mathscr{U}([0,\theta])$. On rappelle que si $X \hookrightarrow \mathscr{U}([0,\theta])$, $\mathbb{E}(X) = \theta/2$ et $V(X) = \theta^2/12$.
On pose $T_n = 2\overline{X_n}$; on a alors $\mathbb{E}(T_n) = \theta$ et $V(T_n) = \theta^2/3n$.

1. Soit $\varepsilon > 0$. Montrer que $\mathbb{P}(|2\overline{X_n} - \theta| \leqslant \varepsilon) \geqslant 1 - \dfrac{\theta^2}{3n\varepsilon^2}$.

   Pour viser un intervalle au niveau $1-\alpha$ *on pose alors* $\varepsilon = \dfrac{\theta}{\sqrt{3n\alpha}}$.

2. Montrer que
$$\left(2\overline{X_n} - \frac{\theta}{\sqrt{3n\alpha}} \leqslant \theta \leqslant 2\overline{X_n} + \frac{\theta}{\sqrt{3n\alpha}}\right) \iff \left(\frac{2\overline{X_n}}{1+\frac{1}{\sqrt{3n\alpha}}} \leqslant \theta \leqslant \frac{2\overline{X_n}}{1-\frac{1}{\sqrt{3n\alpha}}}\right)$$

3. Montrer que la largeur ℓ_n de cet intervalle vérifie : $\dfrac{\ell_n}{\overline{X_n}} \underset{n\to+\infty}{\sim} \dfrac{K}{\sqrt{n}}$.

Exercice 9. Intervalle de confiance de $\mathscr{U}([0,\theta])$ avec le maximum.

On pose $(X_1,\dots,X_n)$ un échantillon de $\mathscr{U}([0,\theta])$. On définit l'estimateur de θ :
$$M_n = \max(X_1,\dots,X_n)$$
On a vu dans un exercice précédent que $\mathbb{E}(M_n) = \dfrac{n\theta}{n+1}$ et $V(M_n) = \dfrac{n\theta^2}{(n+1)^2(n+2)}$.

1. Montrer que
$$\mathbb{P}((M_n-\theta)^2 > \varepsilon^2) \leqslant \frac{\mathbb{E}((M_n-\theta)^2)}{\varepsilon^2}$$

2. En déduire :
$$\mathbb{P}(|M_n-\theta| > \varepsilon) \leqslant \frac{\theta^2}{\varepsilon^2} f(n)$$
où $f(n) = \dfrac{2}{(n+1)(n+2)}$.

3. On cherche à construire un intervalle de confiance au niveau $1-\alpha$.
   (a) Montrer que pour $\varepsilon = \theta\sqrt{\dfrac{f(n)}{\alpha}}$ on a $\mathbb{P}(|M_n-\theta| \leqslant \varepsilon) \geqslant 1-\alpha$.
   (b) Montrer alors que
   $$\left[\frac{M_n}{1+\sqrt{\frac{f(n)}{\alpha}}}, \frac{M_n}{1-\sqrt{\frac{f(n)}{\alpha}}}\right]$$
   est un intervalle de confiance pour θ au niveau de confiance $1-\alpha$.
   (c) Montrer que la largeur ℓ_n de cet intervalle vérifie : $\dfrac{\ell_n}{M_n} \underset{n\to+\infty}{\sim} \dfrac{K}{n}$.

Exercice 10. (HEC 2015) Soit $(X_n)_{n\in\mathbb{N}^*}$ un échantillon de la loi $\mathscr{E}(\lambda)$. On pose $\overline{X_n}$ les moyennes empiriques des X_i. On sait alors que $\mathbb{E}\left(\overline{X_n}\right) = \dfrac{1}{\lambda}$ et $V\left(\overline{X_n}\right) = \dfrac{1}{n\lambda^2}$.

1. Construire les moyennes centrées réduites $\overline{X_n}^*$.
2. Soit $t_1 \geqslant 0$. Montrer que $\lim\limits_{n\to+\infty} \mathbb{P}(-t_1 \leqslant \overline{X_n}^* \leqslant t_1) = 2\Phi(t_1) - 1$.
3. Montrer que $\left[\dfrac{\overline{X_n}}{1+\frac{t_1}{\sqrt{n}}}, \dfrac{\overline{X_n}}{1-\frac{t_1}{\sqrt{n}}}\right]$ est un intervalle de confiance asymptotique au niveau de confiance $2\Phi(t_1)-1$.
4. Que doit valoir $\Phi(t_1)$ pour un intervalle à 99%? En déduire $t_1 \simeq 2{,}57$ par lecture de la table de Φ.

(NB : on peut alors remplacer t_1 par $t_\alpha = \Phi^{-1}\left(1-\dfrac{\alpha}{2}\right)$ pour obtenir un intervalle de confiance au niveau de confiance $1-\alpha$)

Exercice 11. Soit X une variable aléatoire d'espérance m et de variance v ; soit $(X_1,\dots,X_n)$ un n-échantillon de X. On appelle *variance empirique* de X la variable $W_n = \dfrac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X_n}\right)^2$, où on a noté $\overline{X_n}$ la moyenne empirique des X_i.

1. Montrer que $W_n = \left(\dfrac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \overline{X_n}^2$.
2. On rappelle (mais il faut savoir le redémontrer !!) que $\mathbb{E}\left(\overline{X_n}\right) = m$ et $V\left(\overline{X_n}\right) = \dfrac{v}{n}$. En déduire $\mathbb{E}\left(\overline{X_n}^2\right)$.
3. Calculer $\mathbb{E}(W_n)$. En déduire que $\dfrac{n}{n-1}W_n$ est un estimateur sans biais de v.

Exercice 12 (Les tanks allemands)**.**

Pendant la seconde guerre mondiale, les alliés cherchent à estimer le nombre de tanks allemands, à partir des numéros de série des tanks capturés. Ils supposent que les tanks sont numérotés de 1 à N, et qu'ils ont une chance égale de capturer chacun d'entre eux. On se retrouve donc face à un problème d'estimation traduisible en urnes / boules.
Soit une urne contenant N boules. On tire k boules sans remise ; on souhaite donner une estimation de N à partir du k-échantillon ainsi obtenu. On note M_k la variable aléatoire égale à la valeur maximale des boules tirées.

1. Donner le nombre de parties à k éléments de $[\![1,N]\!]$ dont le plus grand élément vaut m. En déduire $\mathbb{P}(M_k = m)$.
2. Montrer que $\mathbb{E}(M_k) = \dfrac{1}{\binom{N}{k}}\sum_{m=k}^{N} m\binom{m-1}{k-1}$.
3. On donne, pour $k \leqslant m \leqslant N$:
   - $m\binom{m-1}{k-1} = k\binom{m}{k}$ (démontrez-le chez vous !)
   - $\sum_{m=k}^{N}\binom{m}{k} = \binom{N+1}{k+1}$ (démontrez-le chez vous ! par formule de Pascal + téléscopage).

   Montrer que $\mathbb{E}(M_k) = \dfrac{k}{k+1}(N+1)$.
4. En déduire un estimateur de N, d'espérance égale à N.

Annexe : tableau de valeurs de Φ

Intégrale $\Phi(x)$ de la Loi Normale Centrée Réduite $\mathcal{N}(0;1)$.

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} \, \mathrm{d}t \quad \text{et} \quad \Phi(-x) = 1 - \Phi(x).$$

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	**0.9236**	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Exemple de lecture : $\Phi(1,43) \simeq 0,9236$ (en gras).