Chapitre 21

Intervalles de confiance

Compétences attendues.

✓ Déterminer un intervalle de confiance par l'inégalité de Bienaymé-Tchebychev.

✓ Déterminer un intervalle de confiance asymptotique à partir d'une convergence en loi.

Objectif. S'il existe des critères pour juger des qualités d'un estimateur ponctuel T_n de $g(\theta)$ (biais, convergence, $V_\theta(T_n)+(E_\theta(T_n)-g(\theta))^2$), aucun de ces critères ne permet de garantir que la valeur prise par un estimateur T_n à partir d'un échantillon observé $(x_1,\dots,x_n)$ sera « proche » de la valeur $g(\theta)$ du paramètre à estimer. Ainsi, même si T_n est un « bon » estimateur ($V_\theta(T_n)+(E_\theta(T_n)-g(\theta))^2$ petit), on n'est jamais à l'abri de tomber sur un « mauvais » échantillon qui nous donnerait une mauvaise estimation de $g(\theta)$.

La démarche de l'estimation par intervalle de confiance est de contrôler cette incertitude. Elle consiste à construire, à partir de l'échantillon, un intervalle (le plus petit possible) dans lequel se trouve la valeur exacte de $g(\theta)$ avec une grande probabilité, fixée à l'avance.

Mathieu Mansuy

Professeur en ECG deuxième année spécialité mathématiques approfondies au Lycée Louis Pergaud (Besançon)

Page personnelle : `mathieu-mansuy.fr/`

E-mail : `mansuy.mathieu@hotmail.fr`

1 Intervalles de confiance

Dans tout ce paragraphe :

- $(X_1, \dots, X_n)$ est un échantillon i.i.d. de même loi mère de paramètre inconnu $\theta \in \Theta$,
- pour tout $n \in \mathbb{N}^*$, $U_n = \varphi_n(X_1, \dots, X_n)$ et $V_n = \psi_n(X_1, \dots, X_n)$ sont des estimateurs de $g(\theta)$ tels que $P_\theta(U_n \leq V_n) = 1$ pour tout $\theta \in \Theta$ (U_n est inférieur à V_n P_θ-presque sûrement).

1.1 Définition

Définition.

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

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

$$\forall \theta \in \Theta, \quad P_\theta(U_n \leq g(\theta) \leq V_n) \geq 1 - \alpha.$$

  Le réel α est appelé le *risque*.

- Soit $\omega \in \Omega$. L'intervalle $[U_n(\omega), V_n(\omega)]$ est une *réalisation* de l'intervalle de confiance $[U_n, V_n]$, aussi appelé *intervalle de confiance observé*.

Remarque. Très souvent, on recherche un intervalle de confiance de $g(\theta)$ sous la forme d'un intervalle centré en une estimation ponctuelle de $g(\theta)$.

Remarque. C'est à celui qui réalise l'étude de fixer le niveau de confiance $1 - \alpha$ qu'il souhaite, et donc le risque α de commettre une erreur qu'il accepte. Par exemple pour $\alpha = 0.05$, et si $[u_n, v_n]$ est une réalisation de $[U_n, V_n]$, alors on a

$$u_n \leq g(\theta) \leq v_n$$

avec une probabilité de 95%. Il y a cependant 5% de (mal)chance de tomber sur un « mauvais échantillon » qui nous donnera un intervalle de confiance ne contenant pas $g(\theta)$.

La plupart du temps, c'est ce niveau de risque de 0.05 qui est utilisé, et qui est communément accepté par exemple en sciences humaines. Mais dans des domaines plus sensibles où l'on n'a pas vraiment de droit à l'erreur (aérospatiale, physique nucléaire, etc), on travaille avec des niveaux de risque de 0.01, voir moins.

Le saviez-vous ?

D'abord spécialiste d'astronomie, Pierre Simon Laplace (1749 - 1827) publie de nombreux articles ayant trait aux probabilités. En 1785, il se propose de calculer la population du royaume de France en ne recensant qu'une petite partie de la population. Appliquant cette méthode, il estime celle-ci à 25 299 417 habitants. Il s'interroge alors sur l'erreur avec une probabilité de 1000 contre 1 de se tromper. Il introduit ce que nous venons d'appeler un intervalle de confiance et démontre la convergence d'une loi binomiale vers la loi normale.

1.2 Intervalle de confiance par l'inégalité de Bienaymé-Tchebychev

On considère une suite (X_n) de variables aléatoires i.i.d. suivant la même loi de Bernoulli de paramètre p inconnu. Rappelons que $\overline{X_n} = \frac{1}{n} \sum_{k=1}^{n} X_k$ est un estimateur sans biais de p.

On explique comment obtenir un intervalle de confiance pour le paramètre p au niveau de confiance $1 - \alpha$ en appliquant l'inégalité de Bienaymé-Tchebychev à $\overline{X_n}$.

Étape 1 : Appliquer l'inégalité de Bienaymé-Tchebychev à $\overline{X_n}$.

Propriété 1 (*Inégalité de Bienaymé* (1796 - 1878) -*Tchebychev* (1821 - 1894))

Soit X une variable définie sur un espace probabilisé $(\Omega, \mathscr{A}, P)$ admettant une variance. On a :

$$\forall \varepsilon > 0, \quad P(|X - E(X)| \geq \varepsilon) \leq \frac{V(X)}{\varepsilon^2}.$$

Étape 2 : Fixer le niveau de risque.

Étape 3 : Expliciter l'intervalle de confiance I.

Vérifions que l'intervalle de confiance I ainsi obtenu répond bien à ce que nous voulons.

On obtient le résultat suivant (dont la démonstration doit être refaite à chaque utilisation).

Propriété 2

Si $(X_1, \ldots, X_n)$ est un n-échantillon i.i.d. de loi de Bernoulli de paramètre p inconnu, alors

$$\left[\overline{X_n} - \frac{1}{2\sqrt{n\alpha}}, \overline{X_n} + \frac{1}{2\sqrt{n\alpha}}\right]$$

est un intervalle de confiance de p au niveau de confiance $1 - \alpha$.

Remarque. Nous pouvons observer sur cet intervalle deux résultats intuitifs :

- plus le niveau de risque souhaité est petit et plus l'intervalle de confiance est grand ;
- plus l'échantillon est de taille n importante et plus l'intervalle de confiance est petit.

Applications numériques.

- Prenons pour risque $\alpha = 0.05$ et pour taille de notre échantillon $n = 100$. Alors un intervalle de confiance de p au niveau de confiance 0.95 est
$$\left[\overline{X_n} - 0.22, \overline{X_n} + 0.22\right].$$
Notons que l'amplitude de cet intervalle est énorme : 0.44 alors que $p \in [0, 1]$.
- Pour $\alpha = 0.05$ et $n = 1000$ (taille de l'échantillon généralement utilisé par les instituts de sondage), l'intervalle de confiance de p au niveau de confiance 0.95 est
$$\left[\overline{X_n} - 0.07, \overline{X_n} + 0.07\right].$$
- Prenons la démarche inverse : on souhaite estimer p avec une erreur d'au plus 0.01 et à un niveau de risque $\alpha = 0.05$. Alors la taille n de notre échantillon doit satisfaire
$$\frac{1}{2\sqrt{n\alpha}} \leq 0.005 \quad \Rightarrow \quad n \geq 50000.$$
Il faut donc un échantillon de taille 50000 (difficile en pratique...).

Simulation. Prenons le cas du deuxième tour d'une élection présidentielle avec deux candidats A et B. Soit p la proportion (inconnue) de personnes interrogées se prononçant pour le candidat A.

```
p = rd.random()
```

On cherche un intervalle de confiance de p au niveau de confiance $1 - \alpha = 0.95$. On sonde pour cela $n = 1000$ personnes.

```
E = rd.binomial(1,p,1000) # 1000-échantillon observé
```

On calcule la moyenne empirique sur cet échantillon observé.

```
Xbar =np.mean(E) ; print(Xbar)
```

On obtient `Xbar = 0.204`. Une réalisation de l'intervalle de confiance de p au niveau de confiance 0.95 est donc $\left[\overline{X_n} - 0.07, \overline{X_n} + 0.07\right] = [0.134, 0.274]$.

Vérifions pour finir si p appartient bien à notre intervalle de confiance (il y a théoriquement 95% de chance que ce soit bien le cas) :

```
print(p)
```

On obtient `p = 0.2113249` qui est bien dans l'intervalle de confiance.

La méthode exposée précédemment peut se généraliser de la manière suivante.

Méthode. Intervalle de confiance par l'inégalité de Bienaymé-Tchebychev.

*Supposons qu'on dispose d'un estimateur T_n **sans biais** de $g(\theta)$ dont on connait (un majorant de) la variance[a] qui **ne fait pas intervenir** $g(\theta)$. Pour déterminer un intervalle de confiance de $g(\theta)$ à l'aide de l'inégalité de Bienaymé-Tchebychev, on procèdera comme suit :*

Étape 1 : Appliquer l'inégalité de Bienaymé-Tchebychev.

On calcule $E(T_n)$ (qui doit valoir $g(\theta)$) et $V(T_n)$ et on applique l'inégalité de Bienaymé-Tchebychev à T_n :

$$\forall \varepsilon > 0, \quad \underbrace{P(|T_n - g(\theta)| \geq \varepsilon)}_{\text{Proba d'être hors de l'I. de C.}} \leq \underbrace{\frac{V(T_n)}{\varepsilon^2}}_{\text{Niveau de risque}} .$$

Étape 2 : Fixer le niveau de confiance.

*On majore si nécessaire $\frac{V(T_n)}{\varepsilon^2}$ par une quantité qui **ne fait pas intervenir** $g(\theta)$. On fixe ensuite ε afin que ce majorant soit égal à α.*

Étape 3 : Expliciter l'intervalle de confiance.

On résout l'inéquation $|T_n - g(\theta)| < \varepsilon$ afin d'isoler $g(\theta)$, et d'obtenir ainsi l'intervalle de confiance.

[a] C'est le cas pour la moyenne empirique $\overline{X_n}$ en tant qu'estimateur de p pour la loi mère $\mathscr{B}(p)$.

Remarque. Si on dispose d'un estimateur T_n **biaisé** de $g(\theta)$, on peut obtenir de manière analogue un intervalle de confiance de $g(\theta)$ en appliquant cette fois l'inégalité de Markov à $(T_n - g(\theta))^2$:

$$P(|T_n - g(\theta)| \geq \varepsilon) = P\left((T_n - g(\theta))^2 \geq \varepsilon^2\right) \leq \frac{E\left((T_n - g(\theta))^2\right)}{\varepsilon^2}.$$

On procède alors en suivant les mêmes étapes que précédemment, en remplaçant $V(T_n)$ par $E\left((T_n - g(\theta))^2\right)$. Il pourra être utile de se souvenir que :

$$\begin{aligned} E\left((T_n - g(\theta))^2\right) &= E\left(T_n^2 - 2g(\theta)T_n + g(\theta)^2\right) \underset{\text{lin. de } E}{=} E(T_n^2) - 2g(\theta)E(T_n) + g(\theta)^2 \\ &= V(T_n) + E(T_n)^2 - 2g(\theta)E(T_n) + g(\theta)^2 = V(T_n) + (E(T_n) - g(\theta))^2. \end{aligned}$$

2 Intervalles de confiance asymptotiques

2.1 Définition

Outre les inégalités de Bienaymé-Tchebychev et de Markov, le théorème central limite permet aussi d'obtenir des estimations par intervalle de confiance. Mais celui-ci donne seulement un résultat asymptotique, d'où la notion suivante.

Définition.

Soit $\alpha \in [0, 1]$. On appelle *intervalle de confiance asymptotique de $g(\theta)$ au niveau de confiance* $1 - \alpha$ toute suite $([U_n, V_n])_{n\in\mathbb{N}^*}$ vérifiant : pour tout $\theta \in \Theta$, il existe une suite de réels $(\alpha_n)_{n\in\mathbb{N}^*}$ à valeurs dans $[0, 1]$ et de limite α, telle que :

$$\forall n \in \mathbb{N}^*, \quad P_\theta(U_n \leq g(\theta) \leq V_n) \geq 1 - \alpha_n.$$

Remarque. Notons la différence avec un intervalle de confiance de niveau $1 - \alpha$: un intervalle de confiance asymptotique sera à un niveau de confiance « acceptable » pour n grand (α_n proche de α), sans plus d'information sur le n à considérer. D'où une perte de précision ici.

2.2 Intervalle de confiance asymptotique du paramètre d'une loi de Bernoulli

On considère une suite (X_n) de variables aléatoires i.i.d. suivant la même loi de Bernoulli de paramètre p inconnu. On explique ici comment obtenir à l'aide du Théorème central limite un intervalle de confiance pour p.

Étape 1 : Énoncer le Théorème central limite.

Théorème 3 (*Théorème central limite*)

Hypothèses :
- Les variables aléatoires $(X_n)_{n\in\mathbb{N}^*}$ sont i.i.d.
- Elles admettent une espérance m et une variance σ^2 non nulle.

Pour tout $n \in \mathbb{N}^*$, on note $\overline{X_n} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$, et $\overline{X_n}^*$ la variable centrée réduite associée :

$$\overline{X_n}^* = \frac{\overline{X_n} - E(\overline{X_n})}{\sqrt{V(\overline{X_n})}} = \sqrt{n}\left(\frac{\overline{X_n} - m}{\sigma}\right).$$

Alors $\overline{X_n}^*$ converge en loi vers une variable aléatoire suivant une loi $\mathcal{N}(0,1)$.

Dans notre cas, on a

$E(\overline{X_n}) =$; $V(\overline{X_n}) =$

On en déduit par le Théorème limite central que

$$\sqrt{n}\left(\frac{\overline{X_n} - p}{\sqrt{p(1-p)}}\right) \xrightarrow{\mathcal{L}} X \quad \text{où} \quad X \hookrightarrow \mathcal{N}(0,1).$$

On obtient pour tous $a < b$ réels :

$$\lim_{n\to+\infty} P\left(a < \sqrt{n}\frac{\overline{X_n} - p}{\sqrt{p(1-p)}} \leq b\right) = P(a < X \leq b) = \Phi(b) - \Phi(a).$$

Étape 2 : Fixer le niveau de confiance.

On souhaite obtenir un intervalle de confiance asymptotique au niveau de confiance $1-\alpha$. Pour cela, on doit avoir :

$$\Phi(b) - \Phi(a) \geq 1 - \alpha.$$

Il y a une infinité de façon de choisir a et b. On choisit couramment $a = -b$ (intervalle centré en la moyenne empirique). Dans ce cas, on a :

$$\Phi(b) - \Phi(a) = \Phi(b) - \Phi(-b) = \Phi(b) - (1 - \Phi(b)) = 2\Phi(b) - 1.$$

On cherche alors b tel que :

$$2\Phi(b) - 1 = 1 - \alpha \quad \Leftrightarrow \quad \Phi(b) = 1 - \frac{\alpha}{2}.$$

Φ étant continue et strictement croissante sur $\mathbb{R}$, elle réalise une bijection de $\mathbb{R}$ dans $\Phi(\mathbb{R}) =]0,1[$. Il existe donc un unique réel t_α tel que $\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$. On pose ainsi pour la suite $b = t_\alpha$, et $a = -t_\alpha$.

Étape 3 : Expliciter l'intervalle de confiance.

On résout l'inéquation suivante afin d'isoler p :

$$\begin{aligned} -t_\alpha < \sqrt{n}\frac{\overline{X_n}-p}{\sqrt{p(1-p)}} \leq t_\alpha \quad &\Leftrightarrow \quad -t_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} < \overline{X_n} - p \leq t_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \\ &\Leftrightarrow \quad \overline{X_n} - t_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} < p \leq \overline{X_n} + t_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \end{aligned}$$

On a donc :

$$\lim_{n\to+\infty} P\left(\overline{X_n} - t_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} < p \leq \overline{X_n} + t_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = 1-\alpha$$

Comme précédemment, nous ne connaissons pas p mais nous savons que $p(1-p) \leq \frac{1}{4}$, de sorte que :

$$\lim_{n\to+\infty} P\left(\overline{X_n} - \frac{t_\alpha}{2\sqrt{n}} \leq p \leq \overline{X_n} + \frac{t_\alpha}{2\sqrt{n}}\right) \geq \lim_{n\to+\infty} P\left(\overline{X_n} - t_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} < p \leq \overline{X_n} + t_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = 1-\alpha.$$

On obtient ainsi le résultat suivant (dont la démonstration est à refaire à chaque utilisation).

Propriété 4

Un intervalle de confiance asymptotique du paramètre p d'une loi de Bernoulli au niveau de confiance $1-\alpha$ est :

$$\left[\overline{X_n} - \frac{t_\alpha}{2\sqrt{n}}, \overline{X_n} + \frac{t_\alpha}{2\sqrt{n}}\right].$$

Remarque. Voici deux valeurs de t_α qu'on rencontrera souvent (et qu'on peut retrouver dans la table de la loi normale si besoin) :

$$t_{0.05} = \Phi^{-1}(0.975) \approx 1.96 \text{ pour } \alpha = 0.05 \quad \text{et} \quad t_{0.01} = \Phi^{-1}(0.995) \approx 2.57 \text{ pour } \alpha = 0.01.$$

Applications numériques.

- Prenons pour risque $\alpha = 0.05$ et une taille raisonnable pour l'échantillon $n = 1000$. On obtient l'intervalle de confiance asymptotique

  $$\left[\overline{X_n} - 0.031, \overline{X_n} + 0.031\right].$$

  Il est d'amplitude 0.062, à comparer à 0.14 obtenue pour celui avec l'inégalité de Bienaymé-Tchebychev.

- Si on souhaite estimer p avec une erreur d'au plus 0.01 et un risque $\alpha = 0.05$, alors la taille n de notre échantillon doit satisfaire

  $$\frac{1.96}{2\sqrt{n}} \leq 0.01 \quad \Leftrightarrow \quad n \geq \left(\frac{0.98}{0.01}\right)^2 = 9604.$$

  On a donc $n = 9604$. Là aussi, c'est bien meilleur que le $n = 50000$ obtenu à l'aide de l'inégalité de Bienaymé-Tchebychev.

Simulation. Reprenons notre simulation. On a obtenu sur notre échantillon de taille $n = 1000$ une moyenne empirique observée égale à `Xbar = 0.204`. On obtient donc l'intervalle de confiance asymptotique de `p` au niveau de risque $\alpha = 0.05$ suivant :

$$\left[\overline{X_n} - 0.031, \overline{X_n} + 0.031\right] = [0.173, 0.235],$$

`p` valant en réalité 0.2113249.

Remarque. Il s'agit d'un intervalle de confiance asymptotique, dont on ne contrôle donc pas le risque (il faut que n soit « grand » pour que $\alpha_n \approx \alpha$, mais « grand » comment ?). En pratique, on considère que c'est bien un intervalle de confiance de risque α dès que $n \geq 30$, $np \geq 5$, $n(1-p) \geq 5$, conditions d'approximation d'une loi binomiale par une loi normale (ce qu'on fait ici).

Exemple. Le premier tour de l'élection présidentielle de 2002.
Quelques jours avant les élections, des sondages réalisés auprès de 1000 personnes donnaient (estimations ponctuelles par la moyenne empirique) :

14.5% d'intentions de vote à Jean-Marie Le Pen et 17% à Lionel Jospin.

Pourtant les scores finaux ont été de

16.83% pour Le Pen et 16,18% pour Jospin.

Comment l'expliquer ?

À l'aide des intervalles de confiance que nous venons d'obtenir, on pouvait affirmer avec une certitude de 95% que

le score final de Le Pen serait entre 11.5% et 17.5%, et celui de Jospin entre 14% et 20%.

L'intersection de ces intervalles étant loin d'être vide, il était douteux de conclure uniquement sur la base d'une estimation ponctuelle.

Remarque. Soit (X_n) une suite de variables aléatoires i.i.d. de loi de Bernoulli de paramètre p inconnu. Soit $\alpha \in]0,1[$. On a obtenu deux intervalles de confiance au niveau de confiance $1-\alpha$:

- grâce à l'inégalité de Bienaymé-Tchebychev :
$$\left[\overline{X_n} - \frac{1}{2\sqrt{n\alpha}}, \overline{X_n} + \frac{1}{2\sqrt{n\alpha}}\right] \tag{BT}$$
- grâce au théorème limite central :
$$\left[\overline{X_n} - \frac{t_\alpha}{2\sqrt{n}}, \overline{X_n} + \frac{t_\alpha}{2\sqrt{n}}\right] \tag{TLC}$$

Reprenons notre simulation `Python`. On voulait un intervalle de confiance de `p` à un niveau de risque $\alpha = 0.05$. Pour cela, on avait créé un échantillon de taille $n = 1000$. On avait obtenu `Xbar = 0.204`. On obtient ainsi les intervalles de confiance suivants :

$$BT \ : \ [0.134, 0.274] \quad ; \quad TLC \ : \ [0.173, 0.235].$$

Et on a bien $p = 0.2113249$ qui appartient à ces deux intervalles de confiance (on avait en théorie 95% de chance que ce soit effectivement le cas).

Répétons cela pour $m = 10000$ échantillons de taille $n = 1000$ à l'aide du programme suivant, en testant si p est dans chacun des m intervalles de confiance obtenus.

```python
p=0.2113249 ; alpha=0.05 ; t=1.96
n=1000 ; m=10000
BT=0 ; TLC=0
for k in range(m):
    Xn = np.mean(rd.binomial(1,p,n))
    if np.abs(Xn-p)<t/(2*np.sqrt(alpha*n)) :
        BT=BT+1
    if np.abs(Xn-p)<t/(2*np.sqrt(n)) :
        TLC=TLC+1
print(100*BT/m, 100*TLC/m)
```

Ce programme renvoie la proportion d'intervalles de confiance contenant p. Il estime ainsi (à l'aide de la méthode de Monte Carlo) le niveau de confiance réel de chaque intervalle. On obtient :

$$BT = 100 \quad ; \quad TLC = 98.42.$$

C'est donc l'intervalle TLC qui répond le mieux à notre problème : il nous donne une meilleur approximation de p, l'intervalle de confiance étant plus petit, et est d'un niveau de confiance estimé plus proche de 0.95.

Voici d'autres résultats pour différentes valeurs de p.

p **réel**	**BT**	**TLC**
0.5238291	100.0	95.04
0.7667777	100.0	97.94
0.1610254	100.0	99.26
0.0131476	100.0	100.0
0.9775233	100.0	100.0
0.2489265	100.0	97.73
0.3863217	100.0	95.57

Tableau récapitulatif de résultats pour $n = 1000$ et $\alpha = 0.05$ (avec $m = 10000$ répétitions).

Méthode. Détermination d'un intervalle de confiance asymptotique.

Pour déterminer un intervalle de confiance asymptotique d'un paramètre $g(\theta)$, on procèdera comme suit :

Étape 1 : Disposer d'une convergence en loi $Y_n \xrightarrow{\mathscr{L}} Y$.

Elle résultera souvent du Théorème limite central. Il faudra dans ce cas l'énoncer (avec les hypothèses !).

Étape 2 : Fixer le niveau de confiance.

Pour tout $a < b$, on a :

$$\lim_{n \to +\infty} P(a < Y_n \leq b) = P(a < Y \leq b) = F_Y(b) - F_Y(a).$$

On choisit a et b de telle sorte que $F_Y(b) - F_Y(a) = 1 - \alpha$.

Étape 3 : Expliciter l'intervalle de confiance.

On résout l'inéquation $a < Y_n \leq b$ afin d'isoler $g(\theta)$, et d'obtenir ainsi l'intervalle de confiance.