Chapitre 21

Intervalles de confiance

Compétences attendues.

- ✓ Déterminer un intervalle de confiance par l'inégalité de Bienaymé-Tchebychev.
- ✓ Déterminer un intervalle de confiance asymptotique par le Théorème Limite Central.

Mathieu Mansuy - Professeur de Mathématiques en ECS2 au Lycée Louis Pergaud (Besançon)
mansuy.mathieu@hotmail.fr

1 Intervalles de confiance

S'il existe des critères pour juger des qualités d'un estimateur ponctuel T_n de $g(\theta)$ (biais, risque quadratique, convergence), 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 (risque quadratique faible), 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.

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.

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. On explique comment obtenir un intervalle de confiance pour le paramètre p au niveau de confiance $1-\alpha$ grâce à l'inégalité de Bienaymé-Tchebychev.

Propriété 1 (*Inégalité de Bienaymé-Tchebychev*)

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

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

On a vu que la moyenne empirique

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

est un estimateur sans biais (et convergent) de p. On applique l'inégalité de Bienaymé-Tchebychev à $\overline{X_n}$.

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 = rand()
```

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

```
-->E = grand(1,1000,"bin",1,p) \\ 1000-echantillon observé
```

On calcule la moyenne empirique sur cet échantillon observé.

```
-->Xbar = mean(E)
Xbar =
   0.204
```

On en déduit que $\left[\overline{X_n} - 0.07, \overline{X_n} + 0.07\right] = [0.134, 0.274]$ est une réalisation de l'intervalle de confiance de p au niveau de confiance 0.95.

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) :

```
-->p
p =
  0.2113249
```

Remarque. Plus généralement, si T_n est un estimateur **sans biais** de $g(\theta)$ dont on connait (un majorant de) la variance, alors on obtient en procédant comme précédemment un intervalle de confiance pour $g(\theta)$.

2 Intervalles de confiance asymptotiques

2.1 Définition

Outre l'inégalité de Bienaymé-Tchebychev, le théorème central limite permet aussi d'obtenir des estimations par intervalles 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.

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 normale centrée réduite $\mathcal{N}(0,1)$.

Dans notre cas, on a

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

On en déduit par le théorème limite central que

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

On obtient pour tout $a < b$ réels,

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

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), et donc

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

On cherche 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 obtient donc

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

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

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

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

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

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) :

$$t_{0.05} = 1.96 \text{ pour un risque } \alpha = 0.05 \quad \text{et} \quad t_{0.01} = 2.57 \text{ pour un risque } \alpha = 0.01.$$

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 au 0.14 obtenu pour celui avec l'inégalité de Bienaymé-Tchebychev.

- Si on souhaite estimer p avec une erreur d'au plus 0.01 avec 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 peut assurer 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.

2.3 Intervalle de confiance asymptotique de l'espérance

Considérons une suite (X_n) de variables aléatoires i.i.d. suivant la même loi d'espérance m et de variance σ^2 toutes deux inconnues. Dans cette section, on présente une méthode générale pour obtenir un intervalle de confiance asymptotique de m.

On note comme d'habitude $\overline{X_n} = \dfrac{X_1 + \cdots + X_n}{n}$ la moyenne empirique. Par application du théorème limite central, on sait que

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

Idée. On pourrait reprendre les calculs de la section précédente à l'identique, avec σ en lieu et place de $\sqrt{p(1-p)}$. Mais comment se « débarrasser » de σ dans nos calculs ? On ne peut pas utiliser la majoration $\sigma \leq \dfrac{1}{2}$, seulement valable pour une loi de Bernoulli...

Méthode. On suppose qu'on dispose d'un estimateur convergent S_n de l'écart-type σ. On a donc :

$$\frac{\sigma}{S_n} \xrightarrow{P} 1.$$

Mais alors par le théorème de Slutsky, on obtient :

$$\sqrt{n}\frac{\overline{X_n} - m}{S_n} = \sqrt{n}\frac{\overline{X_n} - m}{\sigma} \times \frac{\sigma}{S_n} \xrightarrow{\mathscr{L}} X \times 1 = X.$$

Soit $\alpha \in]0,1[$ le risque choisi, et t_α tel que $\Phi(t_\alpha) = 1 - \dfrac{\alpha}{2}$. On obtient :

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

Après calculs (identiques au cas de variables de Bernoulli), on peut conclure que :

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

Nous avons donc prouvé le résultat suivant (à redémontrer à chaque utilisation).

Propriété 5

Soit (X_n) une suite de variables aléatoires i.i.d.

Hypothèses :
- Les X_i admettent une espérance m et une variance $\sigma^2 > 0$;
- On dispose d'un estimateur S_n convergent de σ.

Un intervalle de confiance asymptotique de m au niveau de confiance $1 - \alpha$ est

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

On termine en proposant deux estimateurs convergents de l'écart-type.

Propriété 6 (Écart-type empirique)

Soit (X_n) une suite de variables aléatoires i.i.d. admettant un moment d'ordre 4.
Alors l'écart-type empirique

$$S_n = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X_n})^2}$$

est un estimateur convergent de l'écart-type σ.

La preuve a été faite en TD.

Propriété 7 (Estimateur convergent de l'écart-type d'une loi de Bernoulli)

Soit (X_n) une suite de variables aléatoires i.i.d. de loi de Bernoulli de paramètre p inconnu.
Alors

$$T_n = \sqrt{\overline{X_n}(1 - \overline{X_n})}$$

est un estimateur convergent de l'écart-type $\sqrt{p(1-p)}$.

Preuve.

□

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 trois 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 :

$$\left[\overline{X_n} - \frac{t_\alpha}{2\sqrt{n}}, \overline{X_n} + \frac{t_\alpha}{2\sqrt{n}}\right] \tag{TLC1}$$

ou

$$\left[\overline{X_n} - t_\alpha \frac{\sqrt{\overline{X_n}(1-\overline{X_n})}}{\sqrt{n}}, \overline{X_n} + t_\alpha \frac{\sqrt{\overline{X_n}(1-\overline{X_n})}}{\sqrt{n}}\right]. \tag{TLC2}$$

Reprenons notre simulation `Scilab`. 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 :

$$BT \ : \ [0.134, 0.274] \quad ; \quad TLC1 \ : \ [0.173, 0.235] \quad ; \quad TLC2 \ : \ [0.179, 0.229].$$

On a bien $p = 0.2113249$ qui appartient à ces trois 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.

```
p=0.2113249
n=1000 ; alpha=0.05 ; t=1.96 ; m=10000
BT=0 ; TLC1=0 ; TLC2=0
for k=1:m
    Xn=mean(grand(1,n,"bin",1,p))
    if abs(Xn-p)<t/(2*sqrt(alpha*n)) then BT=BT+1
    end
    if abs(Xn-p)<t/(2*sqrt(n)) then TLC1=TLC1+1
    end
    if abs(Xn-p)<t*sqrt(Xn*(1-Xn)/n) then TLC2=TLC2+1
    end
end
disp("Proportion d'intervalles de type BT contenant p : ") ; disp(100*BT/m)
disp("Proportion d'intervalles de type TLC1 contenant p : ") ; disp(100*TLC1/m)
disp("Proportion d'intervalles de type TLC2 contenant p : ") ; disp(100*TLC2/m)
```

Ce programme estime (à l'aide de la méthode de Monte Carlo) le niveau de confiance réel de chaque intervalle. On obtient :

$$BT = 100 \quad ; \quad TLC1 = 98.42 \quad ; \quad TLC2 = 94.85.$$

C'est donc l'intervalle TLC2 qui répond le mieux à notre problème : il nous donne une meilleur approximation de p, l'intervalle de confiance étant plus petit, et est bien d'un niveau de confiance ≈ 0.95.

Voici d'autres résultats pour différentes valeurs de p.

p **réel**	**BT**	**TLC1**	**TLC2**
0.5238291	100.0	95.04	95.04
0.7667777	100.0	97.94	94.94
0.1610254	100.0	99.26	95.19
0.0131476	100.0	100.0	94.31
0.9775233	100.0	100.0	94.88
0.2489265	100.0	97.73	94.37
0.3863217	100.0	95.57	94.63

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