SCIENCES SORBONNE UNIVERSITÉ

Projet

Abdoul DIALLO, Diaby DRAME

08/04/2022

Table des matières

1. Introduction

2. Intervalles de confiance (non asymptotique)

2.1. Définition

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

2.2 Intervalle de confiance par l'inégalité de Hoeffding

3. Intervalles de confiance asymptotique

3.1. Définition

3.2. Intervalle de confiance asymptotique du paramètre d'une loi de Bernoulli

4. Performances des intervalles (Simulation)

5. Référence

1. Introduction

En statistique, les intervalles de confiance sont utilisés pour décrire le degré d'incertitude associé à une estimation d'échantillon d'un paramètre inconnu de la loi. La démarche consiste à construire à partir de l'échantillon un intervalle de confiance (le plus petit possible) dans lequel se trouve la valeur exacte du paramètre inconnu. Il existe deux types d'intervalles de confiance : les intervalles de confiance asymptotiques et les intervalles de confiance non asymptotique.

Notre étude portera sur l'illustration des performances des différents intervalles de confiance pour le paramètre inconnu θ de la loi de Bernoulli $B(\theta), \theta \in]0,1[$

Quelques définitions et théorèmes:

Soient $(X_n)_{n\geq 1}$ une suite de variables aléatoires et X une variable aléatoire:

- $(X_n)_{n\geq 1}$ converge en loi vers la variable aléatoire X, noté $X_n \xrightarrow[n\to+\infty]{\mathcal{L}} X$

si pour toute fonction continue et bornée ϕ, on a

$$\mathbb{E}[\phi(X_n)] \underset{n\to+\infty}{\longrightarrow} \mathbb{E}[\phi(X)]$$

- $(X_n)_{n\geq 1}$ converge en loi vers la variable aléatoire X si et seuleument si en tout point de continuité x de F_X (fonction de répartiton de X), on a

$$F_{X_n}(x) \underset{n\to+\infty}{\longrightarrow} F_X(x)$$

- $(X_n)_{n\geq 1}$ converge en probabilité vers la variable aléatoire X, noté $X_n \xrightarrow[n\to+\infty]{\mathbb{P}} X$, si

$$\forall \epsilon > 0, \mathbb{P}(|X_n - X| \geq \epsilon) \underset{n\to+\infty}{\longrightarrow} 0$$

- $(X_n)_{n\geq 1}$ converge presque surêment vers la variable aléatoire X, noté $X_n \xrightarrow[n\to+\infty]{p.s} X$, si

$$\mathbb{P}(\lim_{n\to+\infty} X_n = X) = 1$$

- La fonction $T(X_1, ..., X_n) = T(\mathbb{X})$ est appellée statistique, si c'est une fonction mesurable de $(X_1, ..., X_n)$
- $\mathbb{X} = (X_1, ..., X_n)$ est i.i.d si les X_i sont indépendantes et suivent une même loi.

Théorème: Loi des Grands Nombres

Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d admettant un moment d'ordre 1 d'espérance m i.e $m = \mathbb{E}[X_1] < +\infty$, alors

$$\overline{X_n} := \tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i \xrightarrow[n\to+\infty]{\mathbb{P}|p.s} \mathbb{E}[X_1] = m$$

Théorème: Théorème Central Limite

Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d admettant un moment d'ordre 2 i.e $\mathbb{E}[X_1^2] < +\infty$), d'espérance m et de variance $\sigma^2 > 0$.
Alors,

$$\sqrt{n}(\overline{X_n} - m) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

2. Intervalles de confiance (non asymptotique)

2.1. Définition

Soient $(X_1, ..., X_n)$ une suite variabes aléatoires i.i.d de paramètre inconnnu $\theta \in \Theta \subseteq \mathbb{R}^d (d \geq 1)$ et $\alpha \in [0,1]$. On appelle intervalle de confiance non asymptotique de θ au niveau de confiance $1-\alpha$ tout invervalle aléatoire $(\underline{\theta}(\mathbb{X}), \overline{\theta}(\mathbb{X}))$ dont les deux bornes sont des statistiques et tel que pour tout $\theta \in \Theta$,

$$\mathbb{P}(\theta \in (\underline{\theta}(\mathbb{X}), \overline{\theta}(\mathbb{X}))) \geq 1 - \alpha$$

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

Propriété: Bienaymé-Tchebychev

Soit X une variable aléatoire, alors pour tout $\epsilon > 0$, on a

$$\mathbb{P}(|X - \mathbb{E}[X]| > \epsilon) \leq \frac{\mathbb{V}(X)}{\epsilon^2}$$

Dans notre étude $(X_1, ..., X_n)$ est i.i.d de même loi de Bernoulli $B(\theta)$ avec θ inconnu.

On a $\mathbb{E}[X_1] = \theta$, donc par la loi des grands nombres, un estimateur naturel de θ est donné par la moyenne empirique $\overline{X_n}$.

Par ailleurs,

$$\mathbb{E}[\overline{X_n}] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i]$$

Comme les X_i sont de même loi et par linéarité de l'espérance, on a

$$\mathbb{E}[\overline{X_n}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_1] = \mathbb{E}[X_1] = \theta$$

$$\mathbb{V}(\overline{X_n}) = \mathbb{V}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_1) = \frac{\theta(1-\theta)}{n}$$

En appliquant l'inégalité de Bienaymé-Tchebychev, on a

$$\mathbb{P}(|\overline{X_n} - \theta| > \epsilon) \leq \frac{\theta(1-\theta)}{n\epsilon^2}$$

Puis que $\theta(1-\theta) \leq 1/4$ pour tout $\theta \in]0,1[$, on obtient

$$\mathbb{P}(|\overline{X_n} - \theta| > \epsilon) \leq \frac{1}{4n\epsilon^2}$$

Explicitons un intervalle de confiance de niveau $1 - \alpha$ avec $\alpha \in [0,1]$ pour notre paramètre θ:

On a $\forall \epsilon > 0$,

$\mathbb{P}(|\overline{X_n} - \theta| > \epsilon) \leq \frac{1}{4n\epsilon^2} \Rightarrow 1 - \mathbb{P}(|\overline{X_n} - \theta| \leq \epsilon) \leq \frac{1}{4n\epsilon^2} \Rightarrow \mathbb{P}(|\overline{X_n} - \theta| \leq \epsilon) \geq 1 - \frac{1}{4n\epsilon^2}$

Posons $\alpha = \frac{1}{4n\epsilon^2}$ alors $\epsilon = \frac{1}{2\sqrt{n\alpha}}$, donc

$\mathbb{P}(-\frac{1}{2\sqrt{n\alpha}} \leq \overline{X_n} - \theta \leq \frac{1}{2\sqrt{n\alpha}}) \geq 1-\alpha \Rightarrow \mathbb{P}(\overline{X_n} - \frac{1}{2\sqrt{n\alpha}} \leq \theta \leq \overline{X_n} + \frac{1}{2\sqrt{n\alpha}}) \geq 1-\alpha$

Ainsi un intervalle de confiance (non asymptotique) de niveau $1-\alpha$ pour θ est donné par

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

```r
IC_bc <- function(obs,alpha){
n <- length(obs)
m <- mean(obs)
i <- m - 1/(2*sqrt(n*alpha))
s <- m + 1/(2*sqrt(n*alpha))
return (list(inf_bc = i, sup_bc=s))
}

real <- function(size,theta){
  u <- runif(size)
  B <- u < theta
  return (B)
}

obs <- real(1000,0.4)
IC_bc(obs,0.05)
```

```
## $inf_bc
## [1] 0.3412893
##
## $sup_bc
## [1] 0.4827107
```

2.2 Intervalle de confiance par l'inégalité de Hoeffding

Propriété: Hoeffding

Soit $X_1, ..., X_n$ des variables aléatoires indépendantes et bornées, avec $a_i \leq X_i \leq b_i$. Alors $\epsilon > 0$,

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| > \epsilon) \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

avec $S_n = \sum_{i=1}^n X_i$

En se ramenant à notre étude de la suite $(X_1, ..., X_n)$ i.i.d de même loi de Bernoulli $B(\theta)$ avec θ inconnu, on peut expliciter un intervalle de confiance pour θ comme précédemment à l'aide de cette fois-ci de l'inégalité de Hoeffding:

Les X_i étant de même loi, on peut prendre $a_i = a$ et $b_i = b$ $\forall i \in 1, ..., n$. On aura

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| > \epsilon) \leq 2\exp\left(-\frac{2\epsilon^2}{n(b-a)^2}\right)$$

ϵ étant quelconque, prenons $\epsilon = \epsilon n$, donc

$\mathbb{P}(|\frac{S_n}{n} - \mathbb{E}[\frac{S_n}{n}]| > \epsilon) \leq 2\exp\left(-\frac{2\epsilon^2 n^2}{n(b-a)^2}\right) \iff \mathbb{P}(|\overline{X_n} - \mathbb{E}[\overline{X_n}]| > \epsilon) \leq 2\exp\left(-\frac{2\epsilon^2 n}{(b-a)^2}\right)$

Comme nos X_i sont de Bernoulli, alors $a = 0$ et $b = 1$, d'où

$$\mathbb{P}(|\overline{X_n} - \theta| > \epsilon) \leq 2\exp\left(-2\epsilon^2 n\right)$$

En posant $\alpha = 2exp(-2\epsilon^2 n)$ et suivant la même démarche (celle dans l'inégalité de Bienaymé-Tchebychev), nous obtenons un intervalle de confiance (non asymptotique) de niveau $1 - \alpha$ donné par

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

```
IC_hoef <- function(obs,alpha){
n <- length(obs)
m <- mean(obs)
i <- m - sqrt(-log(alpha/2)/(2*n))
s <- m + sqrt(-log(alpha/2)/(2*n))
return (list(inf_H = i, sup_H=s))
}

real <- function(size,theta){
  u <- runif(size)
  B <- u < theta
  return (B)
}

obs <- real(1000,0.4)
IC_hoef(obs,0.05)
```

```
## $inf_H
## [1] 0.3560531
##
## $sup_H
## [1] 0.4419469
```

3. Intervalles de confiance asymptotique

3.1. Définition

Soient $(X_1, ..., X_n)$ une suite variabes aléatoires i.i.d de paramètre inconnu $\theta \in \Theta \subseteq \mathbb{R}^d (d \geq 1)$ et $\alpha \in [0, 1]$. On appelle intervalle de confiance asymptotique de θ au niveau de confiance $1 - \alpha$ tout invervalle aléatoire $(\underline{\theta}(\mathbb{X}), \overline{\theta}(\mathbb{X}))$ dont les deux bornes sont des statistiques et tel que pour tout $\theta \in \Theta$,

$$\lim_{n \to +\infty} \mathbb{P}(\theta \in (\underline{\theta}(X), \overline{\theta}(X))) \geq 1 - \alpha$$

3.2. Intervalle de confiance asymptotique du paramètre d'une loi de Bernoulli

Considérons une suite de variables aléatoires $(X_1, ..., Xn)$ i.i.d de Bernoulli de paramètre θ inconnu.
On explique ici comment obtenir à l'aide du Théorème Central Limite (TCL) un intervalle de confiance pour θ.

Dans notre cas, on a:

$$\mathbb{E}[X_1] = \theta, \mathbb{V}(X_1) = \theta(1 - \theta)$$

Par le TCL

$$\sqrt{n}(\overline{X_n} - \theta) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \theta(1-\theta))$$

donc

$$\frac{\sqrt{n}}{\sqrt{\theta(1-\theta}}(\overline{X_n} - \theta) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

Par définition de la convergence en loi (celle avec les fonctions de répartition), on obtient pour tous a<b réels:

$$\lim_{n\to+\infty} \mathbb{P}(a < \sqrt{n}(\frac{\overline{X_n} - \theta}{\sqrt{\theta(1-\theta)}}) \leq b) = \mathbb{P}(a < Z \leq b) = \Phi(b) - \Phi(a)$$

Avec $Z \sim \mathcal{N}(0,1)$ et Φ la fonction de répartition de Z.

On souhaite obtenir un intervalle de confiance asymptotique de niveau de confiance $1-\alpha$.

Pour cela, on doit avoir: $\Phi(b) - \Phi(a) \geq 1 - \alpha$.

Il y'a une infinité de façons de choisir a et b. Le choix récurrent est a=-b (intervalle centré en la moyenne empirique). Dans ce cas, on a:

$$\Phi(b) - \Phi(a) = \Phi(b) - \Phi(-b)$$

Par symétrie de Z par rapport à 0, on a $\Phi(-b) = 1 - \Phi(b)$, donc

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

On cherche alors b tels que: $2\Phi(b) - 1 = 1 - \alpha \Leftrightarrow \Phi(b) = 1 - \frac{\alpha}{2}$

Φ étant continue et strictement croissante sur $\mathbb{R}$, elle réalise une bijection de $\mathbb{R}$ dans $\Phi(\mathbb{R}) =]0,1[$. Il éxiste donc un unique réel t_α tel que $\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$. On pose par la suite $b = t_\alpha$ et $a = -t_\alpha$.
Ainsi on obtient que $b = t_\alpha = q_{1-\frac{\alpha}{2}}$ et $a = -t_\alpha = -q_{1-\frac{\alpha}{2}}$ avec $q_{1-\frac{\alpha}{2}} = \Phi^{-1}(1-\frac{\alpha}{2})$ le quantile d'ordre $1-\frac{\alpha}{2}$ de Z.

Comme la loi normale centré reduite est symétrique on a $-q_{1-\frac{\alpha}{2}} = q_{\frac{\alpha}{2}}$

Donc ainsi nous avons

$$\lim_{n\to+\infty} \mathbb{P}(a < \sqrt{n}(\frac{\overline{X_n} - \theta}{\sqrt{\theta(1-\theta)}}) \leq b) = \lim_{n\to+\infty} \mathbb{P}(-q_{1-\frac{\alpha}{2}} < \sqrt{n}(\frac{\overline{X_n} - \theta}{\sqrt{\theta(1-\theta)}}) \leq q_{1-\frac{\alpha}{2}}) = \lim_{n\to+\infty} \mathbb{P}\left(|\overline{X_n} - \theta| \leq \frac{\sqrt{\theta(1-\theta)}}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right) =$$

Comme on sait que θ le paramétre de Bernoulli est dans]0,1[donc on: $\theta(1-\theta) \leq \frac{1}{4}$ alors,

$$\lim_{n\to+\infty} \mathbb{P}\left(|\overline{X_n} - \theta| \leq \frac{1}{2\sqrt{n}} q_{1-\frac{\alpha}{2}}\right) \geq 1 - \alpha$$

Un intervalle de confiance asymptotique pour θ de niveau de confiance $1-\alpha$ est :

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

```
IC_Asym <- function(obs,alpha){
n <- length(obs)
m <- mean(obs)
q <- qnorm(1- alpha/2)
i <- m - q/(2*sqrt(n))
s <- m + q/(2*sqrt(n))
```

```r
return (list(inf_A = i, sup_A=s))
}

real <- function(size,theta){
  u <- runif(size)
  B <- u < theta
  return (B)
}

obs <- real(1000,0.4)
IC_Asym(obs,0.05)
```

```
## $inf_A
## [1] 0.3630102
##
## $sup_A
## [1] 0.4249898
```

4. Performances des intervalles (Simulation)

```r
simul <- function(size,theta,M,alpha){
  P_bc <- 0
  P_H <- 0
  P_A <- 0

  for(i in 1:M){
    obs <- real(size,theta)
    IC_bc <- IC_bc(obs,alpha)
    if(IC_bc$inf_bc <= theta && theta <= IC_bc$sup_bc ) { P_bc <- P_bc + 1}

    IC_hoef <- IC_hoef(obs,alpha)
    if(IC_hoef$inf_H <= theta && theta <= IC_hoef$sup_H ) { P_H <- P_H + 1}

    IC_as <- IC_Asym(obs,alpha)
    if(IC_as$inf_A <= theta && theta <= IC_as$sup_A ) { P_A <- P_A + 1}

  }

  return (list(prop_BC = P_bc/M, prop_Hoef=P_H/M, prop_Asym = P_A/M))
}
```

```r
simul(5,0.36,200,0.05)
```

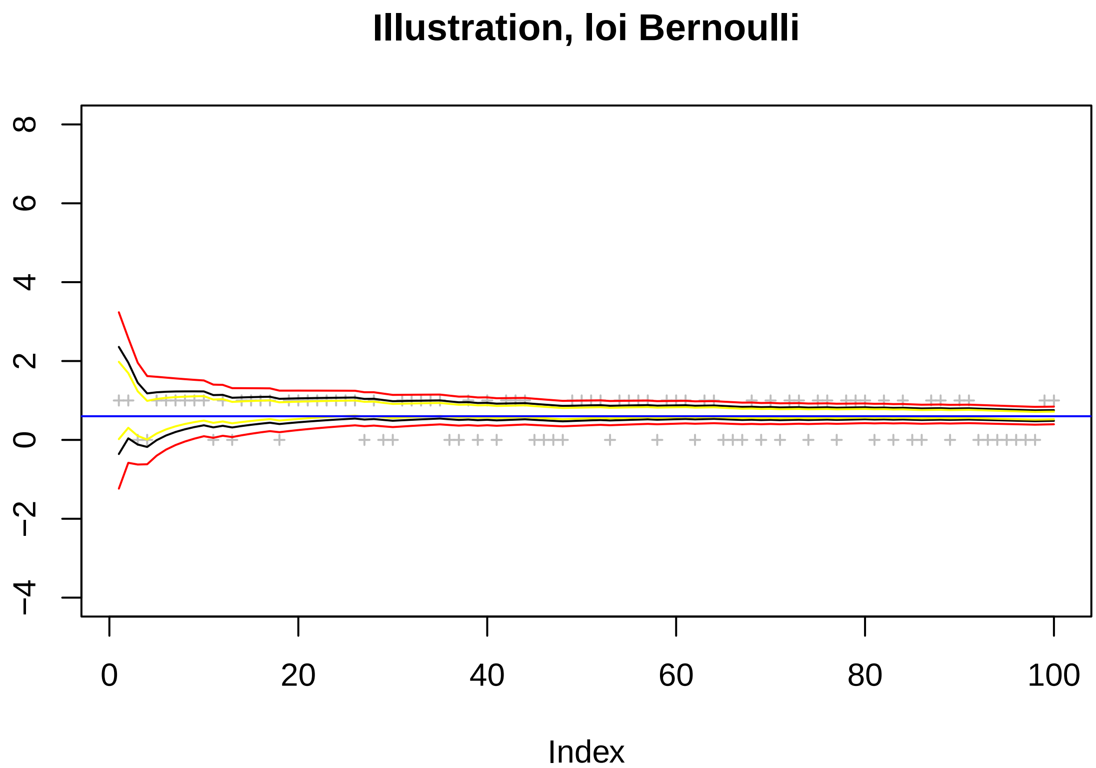
```
## $prop_BC
## [1] 1
##
## $prop_Hoef
## [1] 1
##
## $prop_Asym
## [1] 0.91
```

```r
x = real(100,0.6)
I = NULL
S = NULL
i = NULL
c = NULL

m = NULL
n = NULL

for(k in 1:length(x)){
  a = IC_bc(x[1:k],0.05)
  b = IC_hoef(x[1:k],0.05)
  f = IC_Asym(x[1:k],0.05)
  I[k] = a$inf_bc
  S[k] = a$sup_bc
  i[k] = b$inf_H
  c[k] = b$sup_H
  m[k] = f$inf_A
  n[k] = f$sup_A
}

plot(x, ylim = c(-4,8), ylab = NA, pch = 3, cex=.5,
col="grey", main = "Illustration, loi Bernoulli")
lines(I,col="red")
lines(S,col="red")
lines(i,col="black")
lines(c,col="black")
lines(m,col="yellow")
lines(n,col="yellow")
abline(h=0.6,col="blue")
```



Illustration, loi Bernoulli
8
6
4
2
0
−2
−4
0
20
40
60
80
100
Index

5. Référence

Statistique mathématique, Arnaud Guyader