Principes et Méthodes Statistiques

Valentin Delevoye - Alexandre Frances - Brighton Muffat
Rapport de TP

ENSIMAG 2017-2018

Par soucis de lisibilité, peu de code a été intégré dans le rapport. Veuillez vous référer aux scripts annexes pour voir l'implémentation des fonctions appelées et utilisées dans les différentes questions.

1 Première stratégie

Question 1. Les pêches successives sont indépendantes et à chaque pêche, on remet le poisson péché. La probabilité de pécher un poisson bagué est de $\frac{n_0}{\theta}$. Donc les variables aléatoires $(X_i)_{i\in[|1,n|]}$ (X_i vaut 1 si le i-ème poisson péché est bagué, 0 sinon) sont des variables aléatoires indépendantes et de même loi de Bernoulli de paramètre $\frac{n_0}{\theta}$. Simulons un échantillon de taille n = 100 avec R.

```
theta<-1000
n_0<-50
n<-100
echantillon1<-simulation_echantillon(theta, n_0, n)

>mean(echantillon1) #moyenne empirique
[1] 0.08
>n_0/theta #moyenne theorique
[1] 0.05
>((n-1)/n)*var(echantillon1) #variance empirique
[1] 0.0736
>(n_0/theta)*(1-(n_0/theta)) #variance theorique
[1] 0.0475
```

On observe que la moyenne empirique est supérieure à l'espérance théorique qui vaut 0.05. De même pour la variance empirique et la variance théorique qui vaut 0.0475.

Question 2. T suit une loi binomiale de paramètres $\left(n, \frac{n_0}{\theta}\right)$ en tant que somme de variables aléatoires indépendantes de Bernoulli de paramètre $\frac{n_0}{\theta}$. Donnons sa réalisation t sur l'exemple simulé.

```
>sum(echantillon1)
[1] 8 #realisation de t
```

Question 3.

Estimateur des moments (EMM):

Soit $x_1, ..., x_n$ des réalisations de $X_1, ..., X_n$ identiques et indépendantes de loi $B\left(\frac{n_0}{\theta}\right)$. On cherche à estimer θ.

Puisque $E[X] = \frac{n_0}{\theta} \Rightarrow \theta = \frac{n_0}{E[X]}$, l'estimateur des moments de θ est donc $\tilde{\theta}_n = \frac{n_0}{\overline{X_n}} = \frac{n_0}{\frac{1}{n}\sum_{i=1}^{n} X_i}$.

On obtient donc

$$\boxed{\tilde{\theta_n} = \frac{n\, n_0}{T}} \tag{1}$$

Estimateur de maximum de vraisemblance (EMV) :

Puisque les X_i pour $i \in [\![1, n]\!]$ sont discrètes, de même loi et indépendantes on a :

$$\begin{aligned}
\mathcal{L}\left(\frac{n_0}{\theta}; x_1, ..., x_n\right) &= \prod_{i=1}^{n} P\left(X_i = x_i; \frac{n_0}{\theta}\right) \\
&= \prod_{i=1}^{n} \underbrace{\left(\frac{n_0}{\theta}\right)^{x_i} \left(1 - \frac{n_0}{\theta}\right)^{1-x_i}}_{\| \begin{cases} \frac{n_0}{\theta} & \text{si } x_i = 1 \\ 1 - \frac{n_0}{\theta} & \text{sinon } (x_i = 0) \end{cases}}
\end{aligned}$$

Ainsi on a

$$\begin{aligned}
\mathcal{L}\left(\frac{n_0}{\theta}; x_1, ..., x_n\right) &= \left(\frac{n_0}{\theta}\right)^{x_i} \left(1 - \frac{n_0}{\theta}\right)^{1-x_i} \\
ln\Big(\mathcal{L}\left(\frac{n_0}{\theta}; x_1, ..., x_n\right)\Big) &= \sum_{i=1}^{n} x_i\, ln\Big(\frac{n_0}{\theta}\Big) + \Big(n - \sum_{i=1}^{n} x_i\Big)\, ln\Big(1 - \frac{n_0}{\theta}\Big) \\
\frac{\partial}{\partial\left(\frac{n_0}{\theta}\right)} ln\Big(\mathcal{L}\left(\frac{n_0}{\theta}; x_1, ..., x_n\right)\Big) &= \frac{\Big(\sum_{i=1}^{n} x_i\Big)\,\theta}{n_0} + \frac{\Big(n - \sum_{i=1}^{n} x_i\Big)\,\theta}{\theta - n_0}
\end{aligned}$$

Or cette fonction vaut 0 pour $\frac{n_0}{\theta} = \frac{\sum_{i=1}^{n} x_i}{n}$.

On a donc $\theta = \frac{n\, n_0}{\sum_{i=1}^{n} x_i} = \frac{n\, n_0}{T}$.

Donc l'estimateur de maximum de vraisemblance de θ est

$$\boxed{\widehat{\theta_n} = \frac{n\, n_0}{T} = \tilde{\theta_n}} \tag{2}$$

Ainsi l'estimateur des moments est égal à l'estimateur de maximum de vraisemblance. L'estimation de θ pour l'exemple simulé est donc :

```
>(n*n_0)/t #estimation de l'EMM et l'EMV
[1] 625
```

Question 4. D'après le cours, l'intervalle de confiance exact pour θ est :

$$\boxed{\left[n_0(1+\frac{n-T+1}{T}f_{2(n-T+1),2T,\frac{\alpha}{2}}); n_0(1+\frac{n-T}{T+1}f_{2(n-T),2(T+1),1-\frac{\alpha}{2}})\right]}$$

et l'intervalle de confiance asymptotique est:

$$\boxed{\left[\widehat{p_n} - u_\alpha\sqrt{\frac{\widehat{p_n}(1-\widehat{p_n})}{n}}; \widehat{p_n} + u_\alpha\sqrt{\frac{\widehat{p_n}(1-\widehat{p_n})}{n}}\right]}$$

Avec

$$\boxed{\widehat{p_n} = \frac{n_0}{\widehat{\theta_n}}}$$

Ces intervalles de confiance aux seuils $1\%, 5\%, 10\%$ et 20% pour les données simulées sont :

```
>ic_exact(.01,echantillon1,n_0)
[1] 307.119 1556.778 #IC au seuil 1%
>ic_exact(.05,echantillon1,n_0)
[1] 359.920 1190.941 #IC au seuil 5%
>ic_exact(.10,echantillon1,n_0)
[1] 392.286 1046.975 #IC au seuil 10%
>ic_exact(.20,echantillon1,n_0)
[1] 435.150 908.319 #IC au seuil 20%
```

On a une confiance de 99% dans le fait que θ soit dans l'intervalle trouvé. De même pour les autres intervalles de seuils différents. La valeur 1000 du paramètre θ est comprise dans tous les intervalles sauf celui de seuil 20%

```
>ic_asymptotique(.01,echantillon1,n_0)
[1] 333.599 4940.954 #IC au seuil 1%
>ic_asymptotique(.05,echantillon1,n_0)
[1] 375.452 1863.759 #IC au seuil 5%
>ic_asymptotique(.10,echantillon1,n_0)
[1] 401.207 1413.378 #IC au seuil 10%
>ic_asymptotique(.20,echantillon1,n_0)
[1] 435.662 1105.403 #IC au seuil 20%
```

Asymptotiquement, on a une confiance de 99% dans le fait que θ soit dans l'intervalle trouvé. De même pour les autres intervalles de seuils différents. La valeur 1000 du paramètre θ est comprise dans tous les intervalles.

La largeur de l'intervalle de confiance exact est plus petite que celle de l'intervalle asymptotique. De plus, plus α augmente, plus la largeur de l'intervalle de confiance diminue.

Question 5.

$$P(\widehat{\theta_n} = +\infty) = P(\frac{n_0\, n}{T} = +\infty) \simeq P(T = 0) = \binom{n}{0} \left(\frac{n_0}{\theta}\right)^0 \left(1 - \frac{n_0}{n}\right)^{n-0}$$

$$\boxed{P(\widehat{\theta_n} = +\infty) = \left(1 - \frac{n_0}{n}\right)^n} \tag{3}$$

Puisque $P(\widehat{\theta_n} = +\infty) \neq 0$ si $n_0 < n$, alors $E[\widehat{\theta_n}] \neq \theta$. Donc l'estimateur est biaisé. Cette probabilité dans notre exemple est :

```
>(1-n_0/theta)**n
[1] 0.005920529 # P(estimateur=+inf)
```

Question 6. Après une rapide résolution de l'inéquation suivante :

$$P(\widehat{\theta_n} = +\infty) = \left(1 - \frac{n_0}{n}\right)^n > \frac{1}{2}$$

On trouve

$$\boxed{n \leq \left\lfloor \frac{-ln(2)}{ln(1 - \frac{n_0}{\theta})} \right\rfloor} \tag{4}$$

```
>floor(-log(2)/log(1-n_0/theta))
[1] 13
```

Donc $\boxed{n \in [\![0, 13]\!]}$; Vérifions le expérimentalement en R sur notre exemple :

```
Pour n =  0 P(estimateur = +inf) = 1
Pour n =  1 P(estimateur = +inf) = 0.95
Pour n =  2 P(estimateur = +inf) = 0.9025
Pour n =  3 P(estimateur = +inf) = 0.857375
Pour n =  4 P(estimateur = +inf) = 0.8145062
Pour n =  5 P(estimateur = +inf) = 0.7737809
Pour n =  6 P(estimateur = +inf) = 0.7350919
Pour n =  7 P(estimateur = +inf) = 0.6983373
Pour n =  8 P(estimateur = +inf) = 0.6634204
Pour n =  9 P(estimateur = +inf) = 0.6302494
Pour n =  10 P(estimateur = +inf) = 0.5987369
Pour n =  11 P(estimateur = +inf) = 0.5688001
Pour n =  12 P(estimateur = +inf) = 0.5403601
Pour n =  13 P(estimateur = +inf) = 0.5133421
Pour n =  14 P(estimateur = +inf) = 0.487675
Pour n =  15 P(estimateur = +inf) = 0.4632912
```

L'expérience sur notre exemple met donc en évidence la théorie.

2 Deuxième stratégie

Question 1. Puisque Y_j représente le nombre aléatoire de poissons pêchés entre l'obtention du $(j-1)^{ème}$ et du $j^{ème}$ poisson bagué, alors $\forall k \in \mathbb{N}^*, \forall j \in [\![1,m]\!], \mathbb{P}(Y_j = k) = \mathbb{P}$("pêcher (k-1) poissons non bagués puis un poisson bagué"). Par indépendance des pêches successives on a donc,

$$\forall k \in \mathbb{N}^*, \forall j \in [\![1,m]\!], \mathbb{P}(Y_j = k) = (1 - \frac{n_0}{\theta})^{k-1}(\frac{n_0}{\theta})$$

Les Y_j sont donc identiquement indépendantes de même loi $\mathcal{G}(\frac{n_0}{\theta})$. Simulons un échantillon de taille m = 100 avec R.

```
n_0<-50
theta<-1000
m<-100 #on choisit m = 100
echantillon2<-simulation_echantillon2(theta,n_0,m)

>mean(echantillon2) # moyenne empirique
[1] 20.7
>theta/n_0 # moyenne theorique
[1] 20
>((m-1)/m)*var(echantillon2) # variance empirique
[1] 312.81
>(1-(n_0/theta))/(n_0/theta)**2 #variance theorique
[1] 380
```

On observe que la moyenne empirique est supérieure à l'espérance théorique qui vaut 20, et que la variance empirique est inférieure à la variance théorique qui vaut 380.

Question 2.

$N = \sum_{j=1}^{m} Y_j$
Utilisons les fonctions caractéristiques,

$$\begin{aligned}\phi_N(t) &= \mathbb{E}\left[e^{itN}\right] \\ &= \mathbb{E}\left[\prod_{j=1}^{m} e^{itY_j}\right]\end{aligned}$$

Par indépendance des Y_j puis par lemme des coalitions on a,

$$\phi_N(t) = \prod_{j=1}^{m} \mathbb{E}\left[e^{itY_j}\right]$$

Puisque les Y_j sont de même loi

$$\begin{aligned}\phi_N(t) &= \mathbb{E}\left[e^{itY}\right]^m \\ &= \left(\frac{\frac{n_0}{\theta}.e^{it}}{1-(1-\frac{n_0}{\theta}).e^{it}}\right)^m \\ &= \phi_{\mathcal{NB}(m,\frac{n_0}{\theta})}(t)\end{aligned}$$

Donc $\boxed{N \sim \mathcal{NB}(m, \frac{n_0}{\theta})}$. Donnons sa réalisation n sur l'exemple simulé.

```
>sum(echantillon2)
[1] 2070 # realisation de n
```

Question 3.

Estimateur des moments (EMM) :

Soit $y_1, ..., y_m$ des réalisations de $Y_1, ..., Y_m$ identiques et indépendantes de loi $\mathcal{G}(\frac{n_0}{\theta})$. On cherche à estimer θ.

$\mathbb{E}[Y] = \dfrac{\theta}{n_0}$ donc $\theta = n_0\mathbb{E}[Y]$

Ainsi, $\widetilde{\theta}'_m = n_0\overline{Y_m} = \dfrac{n_0}{n}\displaystyle\sum_{j=1}^{m} Y_j = \dfrac{n_0 N}{n}$

$$\boxed{\widetilde{\theta}'_m = \frac{n_0 N}{n}} \tag{5}$$

Estimateur de maximum de vraisemblance (EMV) :

Puisque les Y_j sont discrètes, identiques et indépendantes on a :

$$\begin{aligned}\mathcal{L}\left(\frac{n_0}{\theta}; y_1, ..., y_m\right) &= \prod_{j=1}^{m} P\left(Y = y_j; \frac{n_0}{\theta}\right) \\ &= \prod_{j=1}^{m}\left(\frac{n_0}{\theta}\right)\left(1-\frac{n_0}{\theta}\right)^{y_j-1} \\ &= \left(\frac{n_0}{\theta}\right)^m \prod_{j=1}^{m}\left(1-\frac{n_0}{\theta}\right)^{y_j-1}\end{aligned}$$

Ainsi on a

$$ln\Big(\mathcal{L}\left(\frac{n_0}{\theta};y_1,...,y_n\right)\Big) = m\,ln\Big(\frac{n_0}{\theta}\Big) + \sum_{j=0}^{m}(y_j-1)\,ln\Big(1-\frac{n_0}{\theta}\Big)$$

$$ln\Big(\mathcal{L}\left(\frac{n_0}{\theta};y_1,...,y_n\right)\Big) = m\,ln\Big(\frac{n_0}{\theta}\Big) + \ ln\Big(1-\frac{n_0}{\theta}\Big)\left(\sum_{j=0}^{m}(y_j)-m\right)$$

$$\frac{\partial}{\partial\left(\frac{n_0}{\theta}\right)}ln\Big(\mathcal{L}\left(\frac{n_0}{\theta};y_1,...,y_n\right)\Big) = \frac{m\theta}{n_0}+\frac{1}{1-\frac{n_0}{\theta}}\left(\sum_{j=1}^{m}y_j-m\right) \text{ s'annule pour } \frac{n_0}{\theta}=\frac{m}{\sum\limits_j y_j}$$

Donc

$$\boxed{\widehat{\theta}'_m = \frac{n_0}{m}N = \widetilde{\theta}'_m} \tag{6}$$

Ainsi l'estimateur des moments est égal à l'estimateur de maximum de vraisemblance. L'estimation de θ pour l'exemple simulé est donc :

```
>(n_0*n)/m
[1] 1035 # estimation de l'EMV et de l'EMM
```

Question 4.

$$\begin{aligned}
I_m(\theta) &= Var\left[\frac{\partial}{\partial\theta}ln\Big(\mathcal{L}\left(\frac{n_0}{\theta},y_1,...,y_n\right)\Big)\right]\\
&= Var\left[\frac{\partial ln\Big(\mathcal{L}\left(\frac{n_0}{\theta},y_1,...,y_n\right)\Big)}{\partial\frac{n_0}{\theta}}.\frac{\partial\frac{n_0}{\theta}}{\partial\theta}\right]\\
&= Var\left[\frac{-n_0}{\theta^2}\left(\frac{m\theta}{n_0}+\frac{1}{1-\frac{n_0}{\theta}}\left(\sum_{j=1}^{m}y_j-m\right)\right)\right]\\
&= Var\left[-\frac{m}{\theta}-\frac{n_0}{\theta^2-\theta n_0}\left(\sum_{j=1}^{m}y_j-m\right)\right]\\
&= \frac{n_0^2}{\theta^2(\theta-n_0)^2}\sum_{j=1}^{m}Var[Y_j]\\
&= \frac{n_0^2.m}{\theta^2(\theta-n_0)^2}\frac{1-\frac{n_0}{\theta}}{\left(\frac{n_0}{\theta}\right)^2}\\
&= \frac{n_0^2.m}{\theta^2(\theta-n_0)^2}\frac{(\theta-n_0)\theta^2}{\theta n_0^2}\\
&= \frac{m}{(\theta-n_0)\theta}
\end{aligned}$$

On a donc

$$\boxed{I_m(\theta) = \frac{m}{(\theta-n_0)\theta}} \tag{7}$$

Par ailleurs,

$$\mathbb{E}\left[\widehat{\theta}'_m\right] = \mathbb{E}\left[\frac{n_0}{m}\sum_{j=1}^{m} Yj\right]$$

Par linéarité, on a

$$\mathbb{E}\left[\widehat{\theta}'_m\right] = \frac{n_0}{m}\sum_{j=1}^{m}\mathbb{E}[Y_j]$$

Puisque les Y_j sont de même loi on a,

$$\begin{aligned}\mathbb{E}\left[\widehat{\theta}'_m\right] &= n_0\mathbb{E}[Y] \\ &= n_0\frac{\theta}{n_0} \\ &= \theta\end{aligned}$$

D'où

$$\boxed{\mathbb{E}\left[\widehat{\theta}'_m\right] = \theta} \tag{8}$$

$\widehat{\theta}'_m$ est donc sans biais. De plus par indépendance des Y_j,

$$\begin{aligned}Var[\widehat{\theta}'_m] &= \frac{n_0^2}{m^2}\sum_{j=1}^{m} Var[Y_j] \\ &= \frac{n_0^2}{m}Var[Y] \\ &= \frac{n_0^2}{m}\frac{(1-\frac{n_0}{\theta})}{\left(\frac{n_0}{\theta}\right)^2} \\ &= \frac{n_0^2}{m}\frac{\theta^2(\theta-n_0)}{n_0^2\theta} \\ &= \frac{(\theta-n_0)\theta}{m}\end{aligned}$$

On a donc

$$\boxed{Var[\widehat{\theta}'_m] = \frac{(\theta-n_0)\theta}{m}}$$

Or

$$\boxed{I_m(\theta) \geq \frac{\left[\frac{\partial}{\partial\theta}\mathbb{E}[\widehat{\theta}'_m]\right]^2}{Var[\widehat{\theta}'_m]} = \frac{1}{\frac{(\theta-n_0)\theta}{m}} = I_m(\theta)} \tag{9}$$

Donc $\widehat{\theta}'_m$ est de variance minimale.

Question 5. D'après l'intervalle de confiance asymptotique donné en énoncé, on a les intervalles de confiance asymptotiques suivants :

```
>ic_asymptotique2(.01)
[1] 774.921 1295.079
>ic_asymptotique2(.05)
[1] 837.104 1232.896
>ic_asymptotique2(.10)
[1] 868.920 1201.079
>ic_asymptotique2(.20)
[1] 905.602 1164.397
```

Asymptotiquement, on a une confiance de 99% dans le fait que θ soit dans l'intervalle trouvé. De même pour les autres intervalles de seuils différents. La valeur 1000 du paramètre θ est comprise dans tous les intervalles.
On constate que plus α augmente plus la largeur de l'intervalle de confiance diminue.

3 Application et comparaison des stratégies

Question 1. Calculons l'estimation de θ et les intervalles de confiance exact et asymptotique de seuil 5% selon la première stratégie en considérant les données du fichier Peche.txt.

```
n_0<-100
n<-1000
echantillon3<-scan('Peche.txt')
t<-sum(echantillon3)

>(n*n_0)/t #estimation de theta
[1] 2857.143

>ic_exact(.05,echantillon3,n_0) #intervalle de confiance exact au
    seuil 5% de theta
[1] 2119.118 3946.499

>ic_asymptotique(.05,echantillon3,n_0)
[1] 2155.61 4235.597 #intervalle de confiance asymptotique au
   seuil 5% de theta
```

Question 2. La fonction R permettant de créer le vecteur $y_1, ..., y_m$ à partir du vecteur $x_1, ..., x_n$ est la suivante :

```
transformation_x_vers_y <- function(vecteur)
```

(Veuillez vous référer aux scripts annexes)

Calculons l'estimation de θ et un intervalle de confiance asymptotique de seuil 5% selon la deuxième stratégie.

```
echantillon4<-transformation_x_vers_y(echantillon3)
m<-length(echantillon4)
```

```
n2<-sum(echantillon4)

>(n_0 * n2) / m #estimation de theta
[1] 2840

>ic_asymptotique2(.05)
[1] 1915.837 3764.163 #intervalle de confiance asymptotique pour
   theta de seuil de 5%
```

Question 3. Vérifions si l'hypothèse d'une loi géométrique est pertinente pour le vecteur $y_1, ..., y_m$.
Soit $k \in \mathbb{N}^*$

$$\begin{aligned} F(k) &= 1 - q^k \\ &= 1 - \left(1 - \frac{n_0}{\theta}\right)^k \\ 1 - F(k) &= \left(1 - \frac{n_0}{\theta}\right)^k \\ ln(1 - F(k)) &= k.ln\left(1 - \frac{n_0}{\theta}\right) \end{aligned}$$

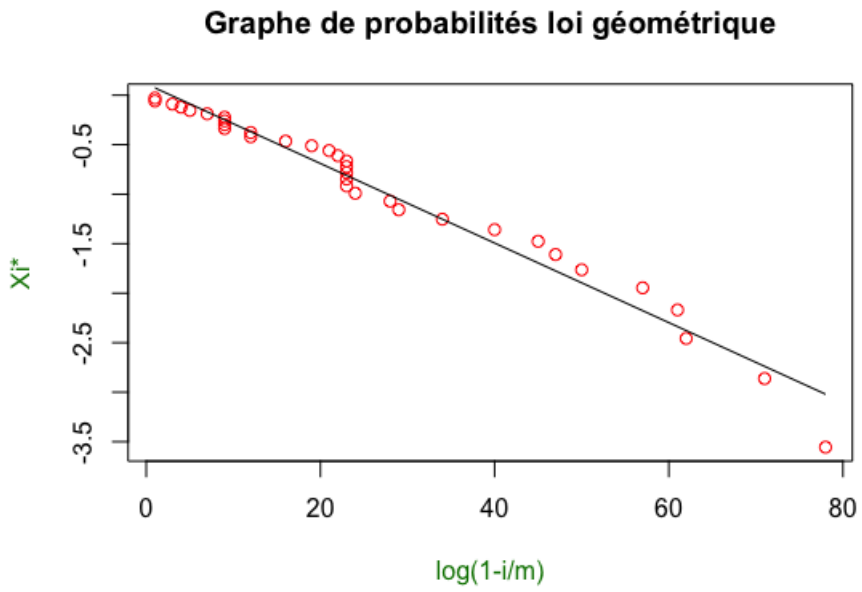
C'est la forme que l'on souhaite avec :

$$\begin{aligned} h(u) &= ln(1-u) \\ \alpha(p) &= ln(1-p) \\ g(k) &= k \\ \beta(p) &= 0 \end{aligned}$$

Le graphe de probabilités pour la loi géométrique est donc le nuage des points $\left(y_i^*; ln\left(1 - \frac{i}{n}\right)\right)_{i \in [\![1, n-1]\!]}$

```
echantillon4ord <- sort(echantillon4)
plot(echantillon4ord[1:(m-1)], log(1-seq(1:(m-1))/m), col='red',
   col.lab=rgb(0,0.5,0), ylab="Xi*", xlab="log(1-i/m)")
title("Graphe de probabilites loi geometrique")

# Superposition de la droite des moindres carres
abs<-echantillon4ord[1:(m-1)]
ord<-log(1 - seq(1:(m-1))/m)
reg<-lm(ord~abs)
lines(abs, fitted.values(reg))
```



On remarque sur cette figure que l'hypothèse de loi géométrique est pertinente.

Question 4. La meilleure stratégie est la deuxième. En effet,
- La largeur de l'intervalle de confiance est plus petite.
- L'hypothèse de loi géométrique est pertinente.
- L'estimateur des moments est un estimateur sans biais de variance minimale, contrairement à celui de la première stratégie.

4 Vérifications expérimentales à base de simulations

Question 1. Simulons m échantillons de taille n de la loi de Bernoulli $\mathcal{B}(\frac{n_0}{\theta})$ et comparons la proportion d'intervalles contenant la vraie valeur de θ à $1-\alpha$

Faisons varier tout d'abord α.

```
>simulation(1000, 50, 100, 1000, .20)
[1] 77,2
>simulation(1000, 50, 100, 1000, .10)
[1] 83.5
>simulation(1000, 50, 100, 1000, .01)
[1] 98.2
```

On remarque que la proportion d'intervalles contenant la vraie valeur de θ est sensiblement égale à $1-\alpha$. Faisons varier θ.

```
>simulation(1000, 50, 100, 1000, .10)
[1] 81.5
>simulation(100, 50, 100, 1000, .10)
[1] 88.6
>simulation(5000, 50, 100, 1000, .10)
[1] 31.9
```

On remarque que au plus theta augmente au plus θ s'éloigne de $1-\alpha$. Cela est dû au fait que la probabilité de pêcher un poisson est trop faible, donc les échantillons ne contiennent quasiment pas de poissons bagués donc il est quasiment impossible de faire de bonnes estimations à moins de pêcher une énorme quantité de poissons et donc d'augmenter n. De plus, les quantiles de la loi de Fisher en R renvoient des "productions de Nan" assez souvent lorsque theta est grand car les nombres sont trop petits ou trop grands. On garde $\theta = 1000$ pour la suite. On conclut de même que le paramètre n_0 influe exactement de manière inverse que le paramètre θ. On garde $n_0 = 50$ pour la suite. Faisons varier n.

```
>simulation(1000, 50, 100, 1000, .10)
[1] 82.1
>simulation(1000, 50, 1000, 1000, .10)
[1] 91.4
>simulation(1000, 50, 10, 1000, .10)
[1] 31.5
>simulation(1000, 50, 10000, 1000, .10)
[1] 90.3
```

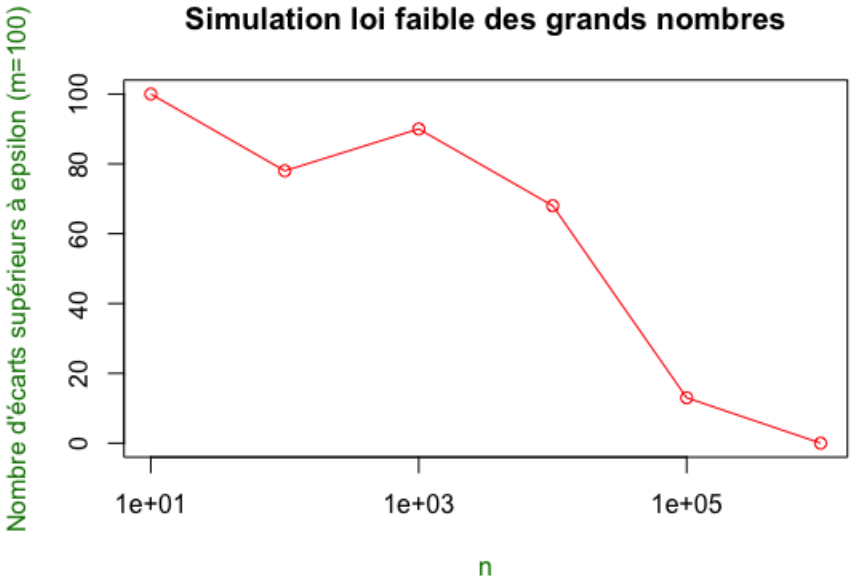
On remarque que plus n augmente, plus la proportion d'intervalles contenant la vraie valeur de θ est proche de $1-\alpha$. Ceci est totalement logique, car plus on a de données plus l'intervalle de confiance est pertinent. Gardons $n = 10000$. Enfin, faisons varier m.

```
>simulation(1000, 50, 10000, 1000, .10)
[1] 90.4
>simulation(1000, 50, 10000, 100, .10)
[1] 91
>simulation(1000, 50, 10000, 10, .10)
[1] 80
>simulation(1000, 50, 10000, 10000, .10)
[1] 90.07
```

On remarque que plus m augmente, plus la proportion d'intervalles contenant la vraie valeur de θ est proche de $1-\alpha$. Ceci est totalement logique, car plus on a de séries statistiques plus l'intervalle de confiance est pertinent.

Question 2. Vérifions la loi faible des grands nombre en faisant la simulation demandée. Faisons les 5 tests suivants que l'on positionne sur un graphe.

```
ord <- c()
abs <- c()
for (i in seq(1:6)) {
    abs <- c(abs, 10**i)
    ord <- c(ord, loi_faible_grds_nb(100, .05, 10**i, .001))
}
plot(abs, ord, log='x', type='o', col="red", xlab='n', col.lab=
   rgb(0,0.5,0), ylab="Nombre d'Ã©carts superieurs a epsilon (m
   =100)", col.lab=rgb(0, 0.5, 0))
title("Simulation loi faible des grands nombres")
```



La loi faible des grands nombres,

$$\forall \varepsilon > 0, \quad \lim_{n \to +\infty} \mathbb{P}\left(\left| \frac{X_1 + X_2 + \cdots + X_n}{n} - E(X) \right| \geqslant \varepsilon \right) = 0$$

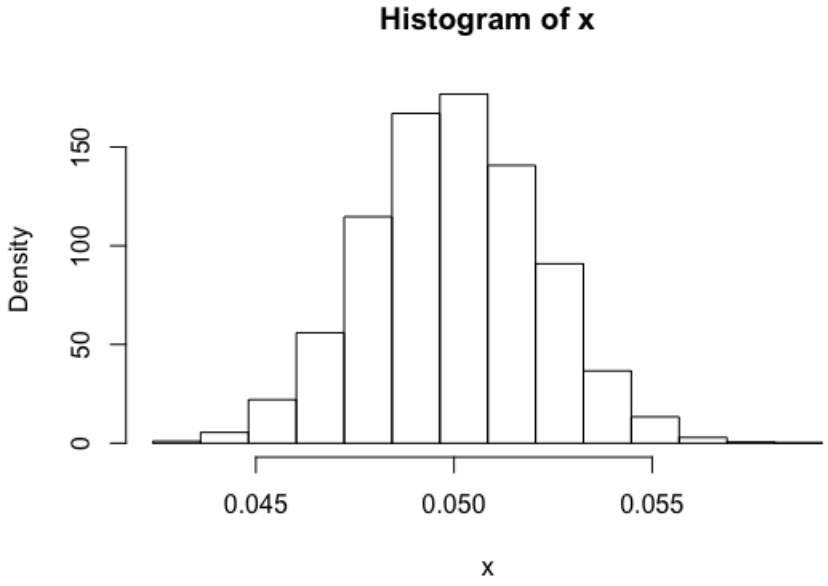
est donc bien vérifiée car sur le graphe lorsque $n \to +\infty$ on a bien l'écart qui tend vers 0.

Question 3. Vérifions le théorème central limite en faisant la simulation demandée.

```
theo_central_limite <- function (m, p, n) {
  vecteur_moyenne <- c()
  for (i in seq(1:m)) {
    echantillon <- rbinom(n, 1, p)
    vecteur_moyenne <- c(vecteur_moyenne, mean(echantillon))
  }
  return(vecteur_moyenne)
}
```

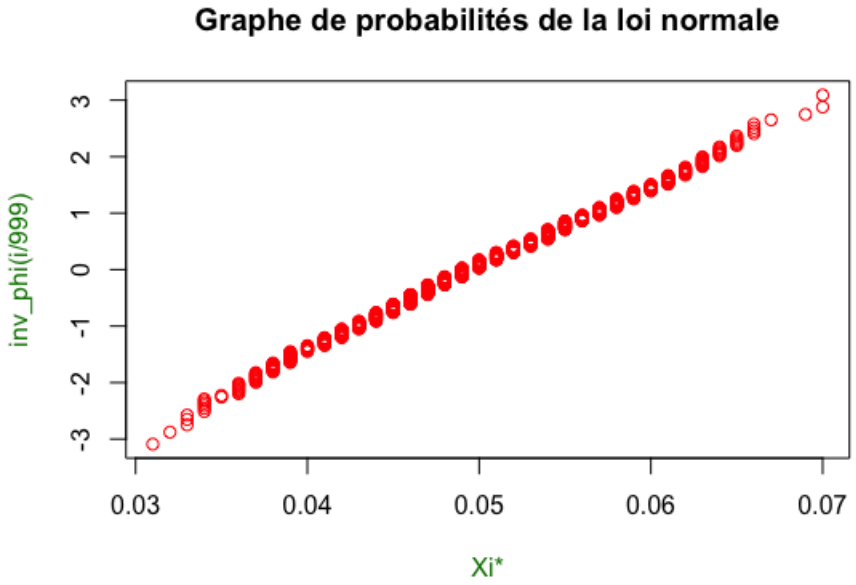
L'histogramme est donc le suivant :

```
histolarg(theo_central_limite(10000, .05, 10000))
```



Et le graphe de probabilités pour une loi normale donne avec ces données simulées :

```
echantillon <- theo_central_limite(1000, .05, 1000)
plot(sort(echantillon)[1:999], qnorm(seq(1:999)/1000), col='red',
    xlab='Xi*', col.lab=rgb(0, 0.5, 0), ylab='inv_phi(i/999)',
   col.lab=rgb(0, 0.5, 0))
title("Graphe de probabilites de la loi normale")
```



Le théorème central limite qui énonce que les moyennes d'un grand nombre d'échantillons suivent une loi normale centrée réduite est vérifié ici. En effet l'histogramme s'apparente bien à une distribution de loi normale avec $n = 10000$ grand et le graphe de probabilités de loi la normale avec les données simulées pour n grand s'apparente bien à une droite.

5 Conclusion

Ce TP nous a permis de découvrir avec curiosité la méthode de capture-marquage-recapture utilisée en écologie et qui donne accès à différents paramètres démographiques et permet une vision plus précise des dynamiques de populations. Ici nous l'avons appliquée à l'estimation du nombre de poissons d'un lac.

En modélisant le problème et en proposant deux différentes stratégies on a pu déterminer la meilleure stratégie. En simulant des estimations du nombre de poissons d'un lac on a pu affiner la stratégie en ajustant les différents paramètres de façon à ce qu'ils soient les plus pertinents possibles et qu'ils permettent d'estimer au mieux le nombre de poissons du lac.

Nous pourrions continuer ce TP en proposant et en étudiant un nouvelle stratégie ou bien en utilisant la stratégie étudiée pour l'appliquer à d'autres problèmes écologiques. En effet, en répétant cette opération d'estimation du nombre de poissons du lac à différents moments, on peut obtenir les variations temporelles des effectifs globaux de la population considérée. Cela nous permettrait d'accéder rapidement à l'état global d'une population : stable, augmentation, diminution, déclin, même si cette méthode n'est pas forcément bien adaptée à tous les écosystèmes.