

Correction du TP sur l'estimation.
Fabienne CASTELL

Exercice 1: Intervalle de confiance pour une proportion.

Soit $(X_i^{(j)}; 1 \leq i \leq n, 1 \leq j \leq N)$ N échantillons de taille n de la loi de Bernoulli de paramètre p . Le programme suivant génère un tel tableau de variables aléatoires (matrice X). Pour chacun des N échantillons, il calcule ensuite les bornes de l'intervalle de confiance de coefficient de sécurité 95 % (matrices Bs et Bi). Ainsi, l'entrée $Bs(i, j)$ correspond à la valeur de la borne supérieure de l'intervalle de confiance pour le j -ième échantillon, stoppé à la i -ème simulation :

$$Bs(i, j) = \frac{1}{i} \sum_{k=1}^i X_k^{(j)} + \frac{1.96}{\sqrt{i}} \sqrt{\frac{\sum_{k=1}^i X_k^{(j)}}{i} \left(1 - \frac{\sum_{k=1}^i X_k^{(j)}}{i}\right)}.$$

On calcule ensuite la proportion de fois où la vraie valeur p n'est pas dans l'intervalle de confiance (vecteur $test$) :

$$\forall i \in \{1, \dots, n\}, \text{test}(i) = \frac{1}{N} \sum_{j=1}^N \mathbb{I}_{p \notin [Bi(i,j); Bs(i,j)]}.$$

Les variables $(\text{test}(i) : 1 \leq i \leq N)$ sont des variables indépendantes, et $N\text{test}(i)$ est une variable binomiale $\mathcal{B}(N, q_i)$, avec

$$q_i \triangleq P \left[\left| \frac{\bar{X}_i}{i} - p \right| \geq \frac{1.96}{\sqrt{i}} \sqrt{\frac{\bar{X}_i}{i} \left(1 - \frac{\bar{X}_i}{i}\right)} \right] \simeq 0.05 \text{ pour } i \text{ grand} .$$

Le programme.

```
% -----
% EstProp
% -----
% Programme relatif a l'exercice I du TP sur l'estimation de proportion.
%

% Dialogue utilisateur. Entree des donnees.
n=input('Nombre de simulations dans un echantillon: ');
N=input('Nombre d''echantillons a simuler: ');
p=input('Proportion a estimer :');

% Simulation des echantillons.
X=rand(n,N);
X=(X <=p);
```

```

% Calcul des moyennes.
M=cumsum(X);
T=[1:n]'*ones(1,N);
M=M./T;
clear X

% Calcul des variances estimees.
S=M.*(1-M);

% Calcul des bornes des l'intervalle de confiance de coefficient 95 %.
niv=0.95;
t=sqrt(2)*erfinv(niv);
E=S./T;
E=sqrt(E);
Bi= M -t * E;
Bs=M+t* E;
clear E t S

% Trace de la vraie valeur et des bornes de l'intervalle sur un
% echantillon.
figure(1)
plot([100:n], [p,p], [100:n]', Bs([100:n],1), 'g', [100:n]', Bi([100:n],1) , 'g')
xlabel('Nombre de simulation')
ylabel(['Bornes de l''intervalle de confiance de niveau ', num2str(niv)])
title('Estimation d''une proportion')

% Calcul du nombre de fois ou la vraie valeur n'est pas dans l'intervalle.
test=(p < Bi) | (p > Bs);
clear Bi Bs
test=mean(test');
figure(2)
plot([100:n],[1-niv 1-niv],[100:n],test([100:n]))
title('estimation du nombre de mauvais echantillons')
xlabel('Taille de l''echantillon')

```

Les résultats.

```

>> rand('state',0)
>> EstProp
Nombre de simulations dans un echantillon: 10000
Nombre d'echantillons a simuler: 1000
Proportion a estimer :0.4

```

Pour l'estimation du nombre de “mauvais” échantillons, on obtient la figure 1. On n'y voit pas vraiment de convergence vers 0.05, mais plutôt des oscillations autour de 0.05. Cela tient au fait qu'on génère des nombres pseudo-aléatoires.

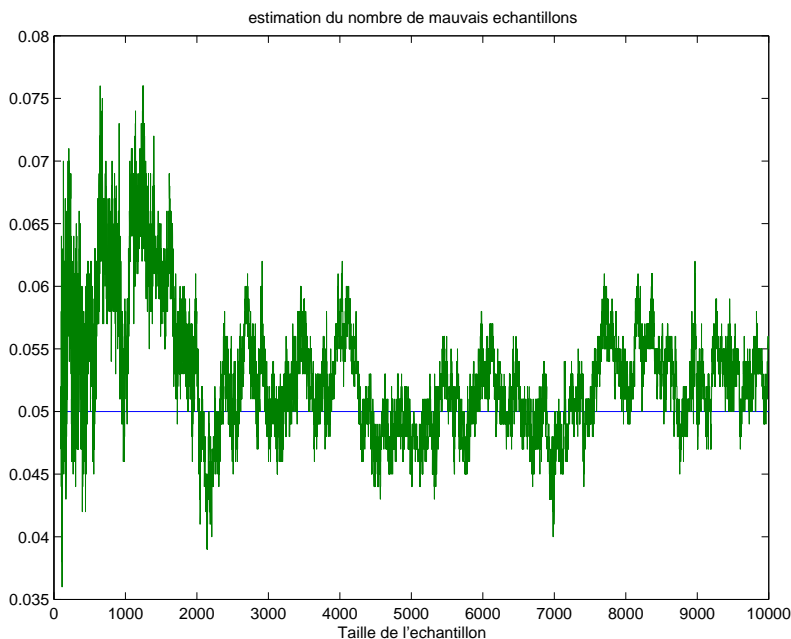


FIG. 1 –

Exercice 2: Intervalle de confiance pour π .

$$N_n = \sum_{i=1}^n \mathbb{I} \{ (X_i, Y_i) \in D \}$$

est une variable binomiale de paramètre n et p

$$p \triangleq P[(X_1, Y_1) \in D] = \int_{-1}^1 \int_{-1}^1 \mathbb{I} \{ x^2 + y^2 \leq 1 \} \frac{dx dy}{4} = \frac{|D|}{4} = \frac{\pi}{4}.$$

On a donc

$$\frac{\sqrt{n}}{\sqrt{\left(\frac{N_n}{n}\left(1 - \frac{N_n}{n}\right)\right)}} \left(\frac{N_n}{n} - \frac{\pi}{4} \right) \xrightarrow[n \rightarrow \infty]{\text{loi}} Z \text{ de loi } \mathcal{N}(0, 1).$$

On en déduit que pour n assez grand

$$\mathbb{P} \left[\frac{4N_n}{n} - 4(1.96) \frac{\sqrt{\frac{N_n}{n}\left(1 - \frac{N_n}{n}\right)}}{\sqrt{n}} \leq \pi \leq \frac{4N_n}{n} + 4(1.96) \frac{\sqrt{\frac{N_n}{n}\left(1 - \frac{N_n}{n}\right)}}{\sqrt{n}} \right] \simeq 0.95.$$

Pour obtenir une approximation de π à 10^{-2} près, on doit donc s'assurer que $4(1.96) \frac{\sqrt{\frac{N_n}{n}\left(1 - \frac{N_n}{n}\right)}}{\sqrt{n}} \leq 10^{-2}$. Comme $x(1-x) \leq 1/4$ pour tout $x \in [0, 1]$, il suffit de choisir n de telle sorte que $4(1.96) \sqrt{\frac{1}{4}} 10^2 \leq \sqrt{n}$, i.e. $n \geq 4 \cdot 10^4 (1.96)^2$.

Le programme suivant met en oeuvre cette méthode d'approximation de π .

Le programme.

```
% EstPi.
%-----
% Programme relatif a l'exercice 2 du Tp sur l'estimation.
```

```

% -----
n=input('Nombre de simulations :');
% Simulation de n uniformes sur le carre [-1;1]^2.
U=2*rand(2,n)-1;

% Test pour savoir si elles sont dans le disque;
U=sum(U.^2);
U=(U<=1);

% Calcul de l'intervalle de confiance pour pi.
U=cumsum(U)./ [1:n]; % calcul de l'estimation de pi/4;
niv=0.95; %coefficient de securite pour l'intervalle de confiance.
tniv=sqrt(2)*erfinv(niv); %quantile correspondant au coefficient de securite.
err=U.*(1-U); % calcul de l'estimation de la variance.
err=err./ [1:n];
err=tniv*sqrt(err);
Bi=4*(U-err); % borne inferieure de l'intervalle de confiance pour pi.
Bs=4*(U+err); % borne superieure de l'intervalle de confiance pour pi.

% representation graphique.
figure(1)
plot([100 n], [pi pi], [100:n], Bi([100:n]), 'g', [100:n], Bs([100:n]), 'g')
title('Estimation de pi')
xlabel('Nombre de simulations.')

```

Les résultats.

```

>> rand('state',0)
>> EstPi
Nombre de simulations :100000

```

On obtient la figure , qui a été zoomée sur les dernières simulations.

Exercice 3: Intervalle de confiance pour des durées de vie.

I/ Cas où on observe tous les τ_i .

La densité du n -uplé (τ_1, \dots, τ_n) par rapport à la mesure $\mu(dx_1, \dots, dx_n) \triangleq \mathbb{1}_{x_1 \geq 0} \cdots \mathbb{1}_{x_n \geq 0} dx_1 \cdots dx_n$ est

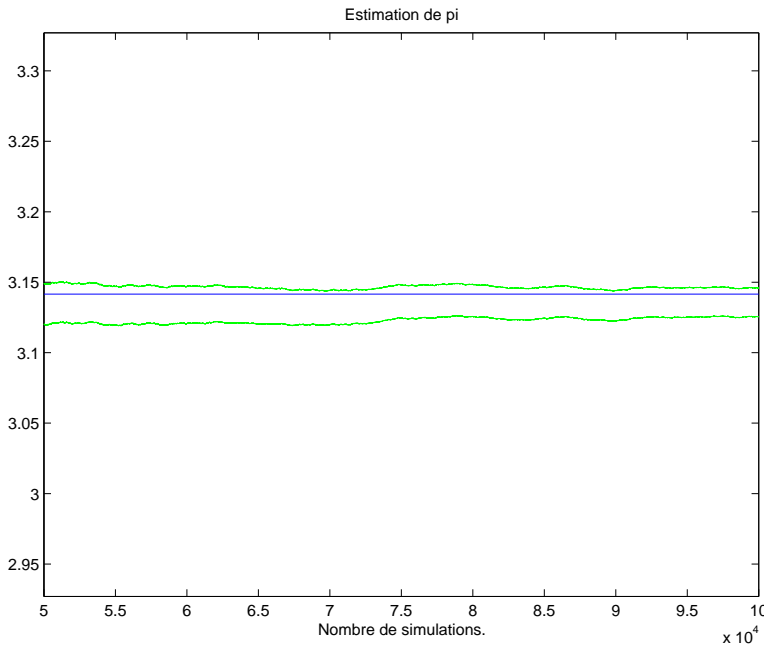
$$f_\theta(x_1, \dots, x_n) = \prod_{i=1}^n \frac{1}{2\theta} \exp\left(-\frac{x_i}{2\theta}\right) = \frac{1}{(2\theta)^n} \exp\left(-\frac{\sum_{i=1}^n x_i}{2\theta}\right).$$

Une vraisemblance du modèle est donc dans ce cas

$$V_\theta(\tau_1, \dots, \tau_n) = \frac{1}{(2\theta)^n} \exp\left(-\frac{\sum_{i=1}^n \tau_i}{2\theta}\right)$$

La maximisation de cette fonction en θ donne pour l'estimateur du maximum de vraisemblance $\hat{\theta}_n = \frac{S_n}{2n}$. On a alors

$$\frac{S_n}{2n\theta} = \frac{1}{n} \sum_{i=1}^n \frac{\tau_i}{2\theta}.$$



Les variables $(\frac{\tau_i}{2\theta}; 1 \leq i \leq n)$ sont des v.a.i.i.d. de loi exponentielle de paramètre 1, et donc la loi de la variable de $\frac{S_n}{2n\theta}$ ne dépend pas de θ . $n\frac{S_n}{2n\theta}$ est en fait la somme de n v.a.i.i.d. de loi $\mathcal{E}(1)$, i.e. une variable de loi $\Gamma(n, 1)$ (densité $\frac{y^{n-1}}{(n-1)!} \exp(-y) \mathbb{1}_{y \geq 0}$). $\frac{S_n}{2n\theta}$ est donc un pivot pour l'estimation de θ . Soit alors $(\underline{t}_n, \bar{t}_n)$ un couple solution de l'équation

$$P(Z \in [\underline{t}_n, \bar{t}_n]) = 95\%, \text{ où } Z \sim \Gamma(n, 1). \quad (1)$$

On a pour tout $\theta > 0$,

$$P_\theta \left[n \frac{S_n}{2n\theta} \in [\underline{t}_n, \bar{t}_n] \right] = 0.95 \Leftrightarrow P_\theta \left[\theta \in \left[\frac{S_n}{2\bar{t}_n}; \frac{S_n}{2\underline{t}_n} \right] \right] = 0.95.$$

Autrement dit $[S_n/(2\bar{t}_n); S_n/(2\underline{t}_n)]$ est un intervalle de confiance pour θ de coefficient de sécurité 95%, dès que $(\underline{t}_n, \bar{t}_n)$ est solution de (1).

L'équation (1) admet une infinité de solution $(\underline{t}_n(\alpha), \bar{t}_n(\alpha))_{\alpha \in [0; 5\%]}$, définie par

$$P[Z \leq \underline{t}_n(\alpha)] = \alpha, \quad P[Z \geq \bar{t}_n(\alpha)] = 5\% - \alpha, \quad Z \sim \Gamma(n, 1).$$

Les valeurs de $\underline{t}_n(\alpha)$ et $\bar{t}_n(\alpha)$ peuvent être obtenues en utilisant la fonction `qgamma` de Stixbox, qui donne la fonction de répartition inverse des lois Gamma.

II/ Cas où on observe seulement les instants des r premières pannes.

Je renvoie au corrigé de l'exercice 16 du chapitre I du cours de stats, pour une démonstration du fait que $\frac{S_r}{2r}$ est l'estimateur du maximum de vraisemblance de θ , et que $\frac{S_r}{\theta}$ suit la loi du χ_r^2 . Remarquez qu'on retrouve les résultats du I/ dans le cas $r = n$ où on observe toutes les pannes. On a en effet dans ce cas, $S_r = T_1 + \dots + T_n = \tau_1 + \dots + \tau_n$, et

$$\begin{aligned} \frac{S_r}{\theta} &\stackrel{\text{loi}}{=} \sum_{i=1}^{2r} X_i^2, \quad X_i \sim \mathcal{N}(0, 1) \text{ indépendantes.} \\ &\stackrel{\text{loi}}{=} \sum_{i=1}^r Z_i, \quad Z_i \sim \mathcal{E}(1/2) \text{ indépendantes, car } \mathcal{E}(1/2) = \chi^2(2) \\ \Leftrightarrow \frac{S_r}{2\theta} &\sim \Gamma(r, 1). \end{aligned}$$

De la même façon que précédemment, si $(\underline{t}_r(\alpha); \bar{t}_r(\alpha))_{\alpha \in [0; 5\%]}$ est solution de

$$P[Z \leq \underline{t}_r(\alpha)] = \alpha, \quad P[Z \geq \bar{t}_r(\alpha)] = 5\% - \alpha, \quad Z \sim \Gamma(r, 1),$$

$[S_r/(2\bar{t}_r(\alpha)); S_r/(2\underline{t}_r(\alpha))]$ est un intervalle de confiance pour θ de coefficient de sécurité 95%.

III/ Simulation numérique.

Le programme suivant commence par simuler un n -échantillon de loi $\mathcal{E}(1/2\theta)$ (n et θ sont rentrés par l'utilisateur). Sur la base de cet échantillon, le programme calcule ensuite les bornes de l'intervalle de confiance lorsqu'on observe les r premières pannes ($1 \leq r \leq n$), pour différentes valeurs de α ($\alpha \in [0, 5\%]$). Si $r \leq 150$, le calcul de $\underline{t}_r(\alpha)$ et $\bar{t}_r(\alpha)$ se fait en utilisant la fonction `qgamma` de `stixbox`. Cette fonction rend des valeurs `NaN` pour des valeurs de $r \geq 170$ (Essayez de comprendre pourquoi...). Aussi pour des valeurs de $r \geq 100$, on utilise pour le calcul de $\underline{t}_r(\alpha)$ et $\bar{t}_r(\alpha)$, une approximation gaussienne. En effet, si $Z \sim \Gamma(r, 1)$, $Z \stackrel{\text{loi}}{=} \sum_{i=1}^r X_i$, où les variables $(X_i; 1 \leq i \leq r)$ sont i.i.d. de loi $\mathcal{E}(1)$ ($E(X_1) = 1$, $\text{var}(X_1) = 1$), et le TLC nous dit que $(Z - n)/\sqrt{n}$ converge en loi quand $n \rightarrow \infty$ vers une variable $\mathcal{N}(0, 1)$.

Le programme.

```
% EstDuree
% -----
% Programme relatif a l'exercice III du TP sur l'estimation.
% -----

% Rentrée des données. Dialogue utilisateur
n = input('Nombre de durees de vie a simuler (>10):');
th= input('Duree moyenne des durees de vie/2 :');

% Simulations des n durees de vie.
X=rand(n,1);
X= - 2*th*log(X);

% Rearrangement par ordre croissant.
X=sort(X);

% Calcul des estimateurs du maximum de vraisemblance.
S=cumsum(X);
S=S+[n-1:-1:0]'.*X;

% Calcul des bornes de l'intervalle de confiance
coef=0.95;

a=[[0.01:0.01:(1-coef)] (1-coef)/2]; %differentes valeurs pour alpha.
a=sort(a);
la=length(a);
% Si n <= 100, on utilise la fonction qgamma. Sinon on utilise une
% approximation gaussienne.
nn=min([n 100]);
NN=[1:nn]'.*ones(1,la);
```

```

A=ones(nn,1)*a;
Bi=qgamma(A,NN);
Bs=qgamma(coef+A,NN);
if n > 100,
    NN=[101:n]';
    Bi=[Bi; NN*ones(1,la)+ sqrt(NN)*qnorm(a)];
    Bs=[Bs; NN*ones(1,la) + sqrt(NN)*qnorm(coef+a)];
end
S=S*ones(1,la);
Bs=S./(2*Bb);
Bi=S./(2*Bi);

% Representation graphique.
deb=min([80,n/2]);

figure(1)
plot(Bi([deb:n],:))
hold on
plot(Bs([deb:n],:))
plot([deb n],[th th],'-r')
axis([0 n-deb th-3 th+3]) %Position des abscisses et des ordonnees
xlabel(['Nombre de pannes observees - ' num2str(deb)])
title('Famille d''IC de coeff 0.95 pour les durees de vie')
legend(['\alpha =' num2str(a(1))], ['\alpha =' num2str(a(2))],...
...['\alpha =' num2str(a(3))],['\alpha =' num2str(a(4))],...
...['\alpha =' num2str(a(5))])
hold off

figure(2)
plot(Bi([deb:n],:)-Bs([deb:n],:))
axis([deb n 0 4])
xlabel(['Nombre de pannes observees - ' num2str(deb)])
title('Longueur de l''IC')
legend(['\alpha =' num2str(a(1))], ['\alpha =' num2str(a(2))],...
...['\alpha =' num2str(a(3))],['\alpha =' num2str(a(4))],...
...['\alpha =' num2str(a(5))])

```

Les résultats.

```

>> rand('state',0)
>> EstDuree
Nombre de durees de vie a simuler (>10):10000
Duree moyenne des durees de vie/2 :5

```

Les figures correspondantes sont les figure 2 et 3. La figure 2 représente les intervalles de confiance pour cinq valeurs de α (0.01, 0.02, 0.025, 0.03, 0.04). Sur la suite de nombres pseudo-aléatoires choisie, la vraie valeur de θ est bien dans l'intervalle de confiance, et cet

intervalle s'améliore quand r augmente. Ceci devrait être moralement vrai, puisque quand r augmente, on dispose de plus d'informations, et l'estimation devrait être plus précise. La figure 3 représente les longueurs des intervalles de confiance pour les différentes valeurs de α . Il semble que pour r de l'ordre de 1000, le choix optimal de α se situe près de 0.03. Essayez de montrer, ou vérifiez numériquement, que lorsque $r \rightarrow \infty$, le choix optimal de α tend vers 0.025

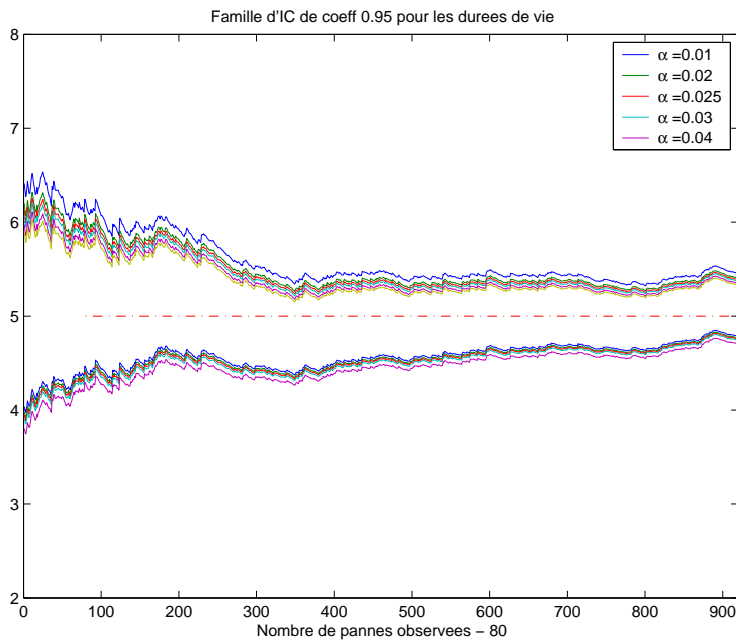


FIG. 2 – Intervalles de confiance pour différents choix de α

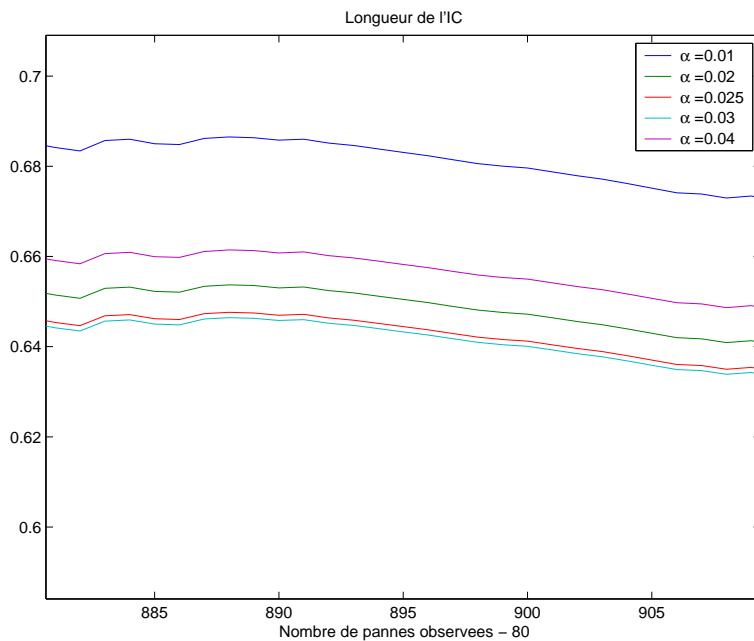


FIG. 3 – Longueur des IC en fonction de α