# TP 3 - Simulation de Monte Carlo et Calcul d'intervalles de confiance

**Introduction :** Soit $X$ un résultat de simulation et ($X_1$ , ... , $X_n$) l'échantillon obtenu en effectuant $n$ réplications d'une simulation. Généralement les calculs d'intervalles de confiance ne sont réalisés que sur des moyennes, l'estimateur utilisé étant la moyenne arithmétique :

$$\overline{X}(n) = \sum_{i=1}^{n} X_i / n$$

Le calcul d'un intervalle de confiance sur cette moyenne arithmétique est trés simple lorsque l'on suppose que les variables $X_i$ ont des distributions identiques, indépendantes et gaussiennes.

**Principe :** L'intervalle de confiance étant centré autour de la moyenne arithmétique, on se borne à calculer le rayon de celui-ci. Le résultat fondamental utilisé pour le calcul de ce rayon est donné ci-après. Si les $X_i$ ont des distributions identiques, indépendantes et gaussiennes de moyenne $\mu$ et de variance $\sigma^2$, alors la variable aléatoire :

$T(n) = \dfrac{\overline{X}(n) - \mu}{\sqrt{S^2(n)/n}}$ est distribuée suivant une loi de Student à *n-1* degrés de libertés, avec :

$S^2(n) = \dfrac{\sum_{i=1}^{n}\left[X_i - \overline{X}(n)\right]^2}{n-1}$ estimateur sans biais de la variance, $\sigma^2$.

Le rayon de l'intervalle de confiance, au niveau de confiance *1-*$\alpha$, est alors donné par:

$$R = t_{n-1,1-\alpha/2} \times \sqrt{\frac{S^2(n)}{n}}$$

où $t_{n-1,1-\alpha/2}$ représente le quantile d'ordre $1-\alpha/2$ d'une loi de Student à *n-1* degrés de libertés

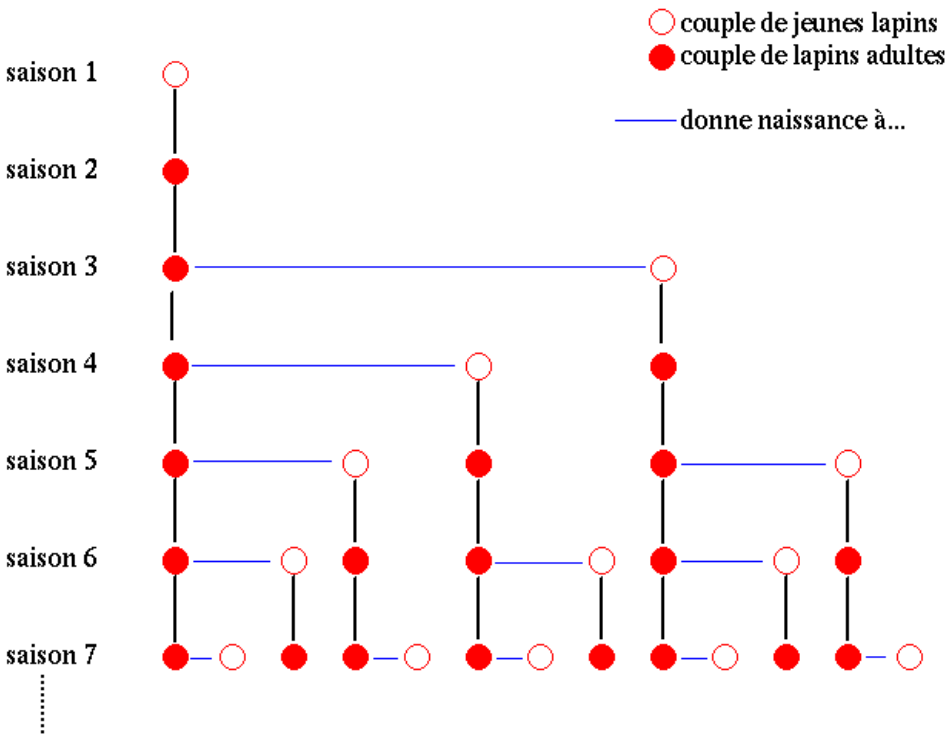
C'est à dire la valeur qui a une probabilité de $1-\alpha/2$ d'être dépassée, ou étant donnée la symétrie de la distribution de Student, une probabilité de 1-$\alpha$ d'être dépassée en valeur absolue). Le tableau 1 donné en annexe donne, en fonction de $n$ les valeurs de $t_{n-1,1-\alpha/2}$ pour $\alpha$=0.05. Le calcul de R donne un intervalle $\left[\overline{X} - R, \overline{X} + R\right]$, au niveau de confiance 1-$\alpha$.

**TP :**

1) Sur le codage simplement en C du calcul de PI avec la méthode de Monte Carlo, faire une boucle pour lancer plusieurs expériences (réplications de ce calcul), en ne réinitialisant pas le générateur de nombres pseudo-aléatoires prendre par exemple le générateur rand() de Linux. Laisser le choix à l'utilisateur quant au nombre de réplications, soit $n$. Calculer un intervalle de confiance à 95% $\alpha$=0.05. Vous pouvez comparer le résultat avec la constante M_PI définie dans <math.h> en fonction du nombre de réplications et d'initialisations différentes du générateur. Attention la comparaison avec M_PI se fait dans la limite du nombre de chiffres significatifs.

2) Simulation de croissance de population : la reproduction des lapins. L'unité de base est un couple de lapins, on considère qu'un couple de jeunes lapins met une saison (1 mois) à devenir adulte, attend une deuxième saison de gestation, puis met au monde un couple de jeunes lapins à chaque saison suivante. En supposant que les lapins ne meurent jamais, on obtient donc le schéma ci-dessous (source :

http://barbara.petit.free.fr/ - vous pourrez consulter ce site et ses liens pour des exemples du nombre d'or). Simuler cette reproduction.




3) Simulation de croissance plus réaliste en C++. Dans la réalité un éleveur de lapins ne pourrait pas compter obtenir exactement un tel rendement: observons par exemple la reproduction des lapins de Garenne. On constate tout d'abord que les mâles sont polygames. Il est donc difficile de raisonner en termes de couples. En suite, Fibonacci avait considéré qu'une période de gestation durait une saison, elle dure en fait 28 à 33 jours. Ceci étant, une femelle met bas 1 à 7 portées par an, mais avec de plus grandes chances que cela soit 3,4 ou 5 portée par an. Quand au nombre de petits dans une portée, il y en a en réalité 3 à 12. Concernant la maturité sexuelle des lapins, elle est en gros de 3 mois pour les femelles, et de 4 pour les mâles. Les lapins ont une durée de vie moyenne de 9 ans. En prenant en compte ces quelques informations et en faisant vos propres choix (adaptation des paramètres, de loi,…), essayer de construire un modèle en proposant un pas de temps, au besoin des histogrammes…
(Ex. : peut-être faut-il considérer une gaussienne autour de 9 ans pour la mortalité, des histogrammes pour les portées et le nombre de petits par portée, etc…).
Etudier ce modèle stochastique avec des réplications.

**Annexes :** Il peut être également intéressant de lui fournir une prévision du nombre de réplications nécessaire pour obtenir une précision à niveau de confiance 1-α fixé:

Nous avons vu que le calcul de $R = t_{n-1,1-\alpha/2}\sqrt{\frac{S^2(n)}{n}}$ donne un intervalle $\left[\overline{X} - R, \overline{X} + R\right]$,

au niveau de confiance 1-α.

Pour obtenir une précision $p = R/\overline{X}$ souhaitée, le nombre $n'$ de réplications est prévisible, si l'on considère l'hypothèse selon laquelle $S^2(n') \cong S^2(n) = S^2$ et $\overline{X}(n') \cong \overline{X}(n) = \overline{X}$, doit respecter l'inégalité $\frac{n'}{\left(t_{n'-1,1-\alpha/2}\right)^2} > \frac{S^2}{\left(p\overline{X}\right)^2}$.

Cette valeur peut donc être trouvée par interpolation sur le tableau 2 donné en annexe.

On en déduit l'algorithme suivant:

*Effectuer $n_0$ réplications*
*$n$ <-- $n_0$*
*Tant que* $\frac{n}{\left(t_{n-1,1-\alpha/2}\right)^2} < \frac{S^2(n)}{\left(p\overline{X}(n)\right)^2}$ *faire*
*Evaluer n' tel que* $\frac{n'}{\left(t_{n'-1,1-\alpha/2}\right)^2} > \frac{S^2(n)}{\left(p\overline{X}(n)\right)^2}$
*Effectuer f*(n'-n) réplications supplémentaires*
*n <-- n + f*(n'-n)*
*Fait*

Remarques : $n_0$ est un nombre initial de réplications, qui doit être suffisamment grand pour faire en sorte que $S^2(n_0)$ et $\overline{X}(n_0)$ soient des estimations valables de la variance et de la moyenne de la variable $X$ (qui représente le résultat de simulation sur lequel on est en train de rechercher un intervalle de confiance). Toutefois $n_0$ ne doit pas être trop grand non plus, car sinon on risque d'avoir effectué des réplications pour rien. Il est difficile de proposer une valeur de $n_0$ valable en général, puisque le nombre de réplications nécessaires est fortement lié à la dispersion de la variable $X$, c'est à dire à la variance de cette loi: Si celle ci est faible on a besoin de peu de réplications, mais si celle ci est forte, il faudra beaucoup de réplications. Or pour connaître cette variance que l'on estime dans l'algorithme par $S^2(n)$, on est bien obligé de faire quelques réplications ...

Pour compenser cette incertitude sur la validité de la valeur $n'$ estimée, on introduit un facteur $f$ compris entre 0 et 1, dont le rôle est d'avancer plus prudemment dans le nombre de réplications: Plutôt que d'effectuer toutes les $n'-n$ réplications supplémentaires, au risque d'avoir fait $n'$ réplications, alors que $n'$ était une mauvaise estimation du nombre de réplications vraiment nécessaire, on choisit de n'en effectuer qu'une proportion $f$. Le choix de $f$ pourra être déterminé par le coût relatif d'une réplication (en terme de temps de calcul par exemple) par rapport à celui de l'algorithme d'estimation du nombre de réplications: Si une réplication coute très cher par rapport à une estimation du nombre de réplications, mieux vaut ne pas faire de réplications inutiles, et donc prendre une valeur de $f$ faible. Dans le cas contraire, il ne sert à rien de faire trop d'estimations, et on peut prendre une valeur de $f$ proche de 1.

**Vérification des hypothèses** : Le calcul présenté ci-dessus est basé sur les trois hypothèses selon lesquelles les variables $X_i$ ont des distributions identiques, indépendantes et gaussiennes:

- la première hypothèse est toujours vérifiée dans le cadre de réplications d'une même simulation;
- la seconde est également assez bien vérifiée dans la pratique, pourvu que l'on ait pris quelques précautions élémentaires en ce qui concerne la génération de nombres aléatoires et que le nombre total de tirages aléatoires pour l'ensemble des réplications ne soit pas "trop grand".

En ce qui concerne la première condition de vérification de l'hypothèse d'indépendance, on doit veiller à ce qu'à chaque réplication le générateur redémarre avec des valeurs initiales (semences) indépendantes. Une méthode préconisée consiste à utiliser les valeurs générées à la fin de la réplication précédente, ce qui est certainement une bonne manière de minimiser les auto-corrélations, puisque le générateur a été conçu pour cela.

La seconde condition impose que le nombre total de tirages aléatoires pour l'ensemble des réplications soit suffisamment petit devant la pseudo-période du générateur, c'est à dire la valeur minimale $k$ pour laquelle la $n^{\text{ième}}$ valeur générée est une fonction déterministe de la $(n+k)^{\text{ième}}$ (souvent la $(n+k)^{\text{ième}}$ peut se déduire de la $n^{\text{ième}}$ par translation).

La troisième hypothèse dépend totalement de la nature de la distribution des $X_i$. Dans beaucoup de cas, les valeurs des $X_i$ sont le résultat d'un processus cumulatif qui fait que, en vertu du théorème central limite, leur distribution converge assez rapidement vers celle d'une loi normale.

Dans la pratique, et par niveau croissant de précaution, on peut :

- tout bonnement accepter l'hypothèse sans la vérifier (ce qui est le cas de trés nombreuses études);
- ou bien vérifier cette hypothèse en utilisant un test de normalité (Chi-2 ou Kolmogoroff). L'inconvénient de ces tests est que l'on ne peut pas connaître le risque d'accepter à tort la normalité de la distribution et donc d'utiliser à tort la procédure proposée pour le calcul de l'intervalle de confiance. Dans le cas où le test accepte la normalité, on accepte donc l'hypothèse sans connaître la probabilité pour qu'elle soit fausse; Toutefois, le résultat du test, qui représente une "distance" entre la distribution empirique des $X_i$ et la distribution théorique de la loi normale (et qui se trouve donc en dessous d'un certain seuil d'acceptabilité) nous fournit un indicateur empirique de cette distance.
- enfin effectuer un calcul d'intervalle de confiance approprié à la distribution particulière des $X_i$ (techniques de bootstrap);

Concrètement on constate que l'hypothèse de normalité a l'avantage d'être très robuste. Ceci signifie que la procédure de calcul d'intervalle de confiance proposée dans le cadre de la normalité des $X_i$ donne toujours des résultats très voisins des résultats obtenus avec d'autres méthodes, même lorsque les $X_i$ ne sont pas distribués normalement, ce qui justifie pourquoi il est assez fréquent d'accepter cette hypothèse de normalité sans la vérifier.

| 1≤n≤10 | $t_{n-1,1-\alpha/2}$ | 11≤n≤20 | $t_{n-1,1-\alpha/2}$ | 21≤n≤30 | $t_{n-1,1-\alpha/2}$ | n>30 | $t_{n-1,1-\alpha/2}$ |
|---|---|---|---|---|---|---|---|
| **1** | **12.706** | **11** | **2.201** | **21** | **2.080** | **40** | **2.021** |
| **2** | **4.303** | **12** | **2.179** | **22** | **2.074** | **80** | **2.000** |
| **3** | **3.182** | **13** | **2.160** | **23** | **2.069** | **120** | **1.980** |
| **4** | **2.776** | **14** | **2.145** | **24** | **2.064** | **+∞** | **1.960** |
| **5** | **2.571** | **15** | **2.131** | **25** | **2.060** | | |
| **6** | **2.447** | **16** | **2.120** | **26** | **2.056** | | |
| **7** | **2.365** | **17** | **2.110** | **27** | **2.052** | | |
| **8** | **2.308** | **18** | **2.101** | **28** | **2.048** | | |
| **9** | **2.262** | **19** | **2.093** | **29** | **2.045** | | |
| **10** | **2.228** | **20** | **2.086** | **30** | **2.042** | | |

Tableau 1

| 1≤n'≤10 | $\frac{n'}{\left(t_{n'-1,1-\alpha/2}\right)^2}$ | 11≤n'≤20 | $\frac{n'}{\left(t_{n'-1,1-\alpha/2}\right)^2}$ | 21≤n'≤30 | $\frac{n'}{\left(t_{n'-1,1-\alpha/2}\right)^2}$ | n'>30 | $\frac{n'}{\left(t_{n'-1,1-\alpha/2}\right)^2}$ |
|---|---|---|---|---|---|---|---|
| **1** | **$6.194\ 10^{-3}$** | **11** | **2.271** | **21** | **4.854** | **40** | **9.973** |
| **2** | **0.108** | **12** | **2.527** | **22** | **5.115** | **80** | **20.000** |
| **3** | **0.296** | **13** | **2.786** | **23** | **5.373** | **120** | **30.609** |
| **4** | **0.519** | **14** | **3.043** | **24** | **5.634** | **n'>120** | **0.260*n'** |
| **5** | **0.756** | **15** | **3.303** | **25** | **5.891** | | |
| **6** | **1.002** | **16** | **3.560** | **26** | **6.151** | | |
| **7** | **1.252** | **17** | **3.818** | **27** | **6.412** | | |
| **8** | **1.502** | **18** | **4.078** | **28** | **6.676** | | |
| **9** | **1.759** | **19** | **4.337** | **29** | **6.934** | | |
| **10** | **2.015** | **20** | **4.596** | **30** | **7.195** | | |

Tableau 2