

## Sujet précédent – Contrôle continu n°1 : Simulation et estimation.

Ce TP noté reprend simplement les bases de ce que nous avons vu jusqu'à présent. Vous pouvez utiliser votre travail des TPs précédents, l'aide de R ainsi qu'internet.

### Consignes

Le TP suivant vous est donné sous la forme d'un fichier .pdf, contenant le sujet, ainsi qu'un notebook qui tient lieu de "fiche réponse" : sur celui-ci, des zones sont laissées libres pour inclure votre code ou vos réponses écrites. A la fin du contrôle, vous devrez :

- Enregistrer votre notebook complété par vos soins, le renommer sous la forme `prenom_nom.Rmd`, et me l'envoyer par mail à l'adresse suivante : `luca.ganassali@inria.fr`.
- Si vous le souhaitez, vous pouvez en plus exporter ce notebook (en .html, ou en .pdf) pour me le rendre dans une version compacte si votre code tourne bien. Ceci n'est pas (du tout) obligatoire.

### Conseils

- Ne perdez pas votre temps à recopier l'énoncé sur le notebook.
- Quand cela vous semble nécessaire, veillez à expliquer votre raisonnement et/ou à commenter votre code. Si votre réponse contient des maths, vous n'êtes pas du tout obligés d'utiliser LaTeX, il suffit d'écrire les choses simplement.
- N'hésitez pas à mettre des titres et des couleurs dans vos graphiques : la clarté et la présentation générale du code joueront dans l'évaluation finale.
- Avant de commencer, prenez le temps de faire un tour du sujet. Pour chaque exercice, un barème indicatif provisoire est donné. Si vous bloquez sur une question, n'hésitez pas à la sauter, il y a la place pour faire beaucoup de choses.
- Dans tout le sujet, on pourra utiliser (si l'on veut) la fonction `which` : elle s'applique à un vecteur  $t$  et renvoie la liste des indices correspondant aux éléments de  $t$  satisfaisant une certaine condition. Par exemple si  $\mathbf{t} = \mathbf{c}(0, -3, 4, 7, -2, 0)$ , alors `which(t>=0)` renvoie `c(1,3,4,6)` : les éléments positifs de  $t$  sont aux positions 1, 3, 4 et 6. De même, `which(t==0)` renvoie `c(1,6)`.

Bon courage !

**Exercice 1 : Loi exponentielle (env. 10 points).** Dans cette partie, nous noterons  $\mathcal{E}(\lambda)$  la loi exponentielle de paramètre  $\lambda > 0$ .

1. Simuler une réalisation d'un échantillon de  $n = 400$  variables de loi  $\mathcal{E}(1/2)$  indépendantes. Calculer et afficher la moyenne, la variance de l'échantillon. Commenter les valeurs calculées.
2. Pour ce même échantillon, afficher sur le même graphique la fonction de répartition empirique et la fonction de répartition théorique de la loi en question. Commenter.
3. On considère  $X$  et  $Y$  deux variables indépendantes de loi  $\mathcal{E}(1/2)$ . On s'intéresse à la loi de  $Z := \frac{X}{X+Y}$ . Simuler plusieurs réalisations de  $X, Y$  et représenter l'histogramme des valeurs de  $Z$ . Quelle semble être la loi de  $Z$  ? Commenter.

4. Une entreprise produit des batteries dont la durée de vie en années est modélisée par une loi  $\mathcal{E}(1/15)$  (attention, le paramètre de la loi a changé !). Deux clients  $A$  et  $B$  achètent chacun une batterie (indépendamment). Le client  $A$  dit au client  $B$  : “j’ai au moins une chance sur quatre que ma batterie tienne au moins 10 ans de plus que la tienne”.

A l’aide d’une simulation, déterminer si le client  $A$  a raison. *Indication : il y a plusieurs façons de faire, la méthode la plus rapide étant celle qui utilise la fonction `which`.*

5. La loi du khi-deux à deux degrés de liberté est la loi de  $N_1^2 + N_2^2$  lorsque  $N_1$  et  $N_2$  sont deux gaussiennes standard indépendantes. A l’aide de simulations, donner une illustration graphique de votre choix montrant que la loi du khi-deux à deux degrés de liberté est en fait la loi  $\mathcal{E}(1/2)$ .

*Indication : pour cette question, la façon d’illustrer ce résultat est laissée libre, il y a plusieurs possibilités, c’est à vous de choisir.*

**Exercice 2 : Estimation du paramètre d’une loi de Poisson (env. 10 points).** On considère un échantillon  $X_1, \dots, X_n$  de variables i.i.d. de loi de Poisson de paramètre  $t > 0$ . On cherche à estimer  $t$ . On définit (comme dans le cours) l’estimateur de la moyenne  $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$  et l’estimateur de la variance sans biais  $\tilde{S}_n^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ .

1. Dans cette question, on ne demande aucune simulation, mais une réponse mathématique. Donner deux estimateurs “naturels” de  $t$  : le premier, qu’on notera  $\hat{t}_1$ , devra utiliser  $\bar{X}_n$ , et le second  $\hat{t}_2$ , devra utiliser  $\tilde{S}_n^2$ .

2. Simuler un échantillon  $X_1, \dots, X_n$  pour  $n = 5000$  et  $t = 3$ . Calculer et afficher  $\hat{t}_1$  et  $\hat{t}_2$  pour cet échantillon.

3. Toujours pour  $t = 3$ , simuler  $N = 7000$  échantillons de taille  $n = 5000$ , et calculer ainsi le risque quadratique (estimé) de  $\hat{t}_1$  et de  $\hat{t}_2$ . Afficher les deux risques. Quel semble être le meilleur des deux estimateurs ?

4. Refaire tourner le code précédent pour  $t = 25$ . Comment évoluent les risques de  $\hat{t}_1$  et de  $\hat{t}_2$  ?

5. On introduit à présent la variable  $s_n$  suivante :

$$s_n := \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_i=0}.$$

(Notez que  $\mathbf{1}_{X_i=0}$  est l’indicatrice de l’évènement ‘ $X_i = 0$ ’).

En utilisant la loi des grands nombres, vers quoi converge (p.s.) la variable  $s_n$  ? En déduire un troisième estimateur  $\hat{t}_3$  de  $t$ , utilisant  $s_n$ .

6. Resimuler un échantillon  $X_1, \dots, X_n$  pour  $n = 5000$  et  $t = 3$ . Calculer et afficher  $\hat{t}_3$  pour cet échantillon. *Indication : il y a plusieurs façons de faire, la méthode la plus rapide étant celle qui utilise la fonction `which`.*

7. Recopier et adapter le code de la question 3. pour qu’il affiche aussi le risque quadratique estimé de  $\hat{t}_3$ . On prendra encore  $N = 7000$  et  $n = 5000$ , mais cette fois-ci on prend pour paramètre  $t = 1$ . Classer les trois estimateurs en fonction de leur qualité.

8. Refaites tourner le code précédent, sans changer  $N$  ni  $n$ , mais pour  $t = 5$ . Le classement est-il le même ? A votre avis, pourquoi ?

9. Refaites tourner le code précédent, sans changer  $N$  ni  $n$ , mais pour  $t = 10$ . Que renvoie le code pour le risque de  $\hat{t}_3$  ? A votre avis, pourquoi ?

**Exercice bonus : Loi de Cauchy et loi des grands nombres.** La *loi de Cauchy* est la loi de densité  $f(x) := \frac{1}{\pi(x^2+1)}$  sur  $\mathbb{R}$ .

1. Calculer la fonction de répartition de cette loi.

2. En déduire une méthode pour simuler  $n = 10000$  variables indépendantes de loi de densité  $f$  en utilisant que des variables de loi uniforme sur  $[0, 1]$ . On notera  $X_1, \dots, X_n$  les variables ainsi simulées. Représenter

$$Y_k := \frac{1}{k} \sum_{i=1}^k X_i$$

en fonction de  $k$  pour  $k$  allant de 0 à  $n$ .

3. Répéter la simulation ci-dessus plusieurs fois. Qu'observe-t-on ? Commenter.