TP 12. Comparaison d'estimateurs dans des modèles de fiabilité

Le but de ce TP est de vous familiariser avec la notion d'*estimation statistique*. Plus particulièrement, nous allons *comparer* la qualité de différentes façons d'estimer des paramètres. Les trois premiers modèles illustrant ce propos sont utilisés en fiabilité pour modéliser des durées de fonctionnement. Vous devrez mener l'étude comparative vous-même sur le quatrième modèle.

# 1 Modèle exponentiel

Le modèle probabiliste de base pour modéliser les durées de fonctionnement est la loi exponentielle. Sa densité est paramétrée par $\lambda > 0$ :

$$f(t) = \lambda e^{-\lambda t} \mathbb{1}_{t \geq 0} \, .$$

Un domaine privilégié de la loi exponentielle est le domaine de la radioactivité (Rutherford et Soddy). Chaque atome radioactif possède une durée de vie qui suit une loi exponentielle. Le paramètre $\lambda$ s'appelle alors la constante de désintégration. La durée de vie moyenne $\frac{1}{\lambda}$ s'appelle le temps caractéristique. On modélise aussi fréquemment la durée de vie d'un composant électronique par une loi exponentielle.

## 1.1 Comparaison d'estimateurs dans le modèle exponentiel

Soit $(X_1, ..., X_n)$ un échantillon de la loi exponentielle de paramètre $\lambda$ où $\lambda$ est un paramètre inconnu. On note $(X_{(1)}, ..., X_{(n)})$ les statistiques d'ordre de l'échantillon. Elles vérifient :

$$X_{(1)} \leq ... \leq X_{(n)} \, .$$

On considère les estimateurs suivants du paramètre $\lambda$.

$$T_1 = \frac{n}{\sum_{i=1}^n X_i}$$

$$T_2 = \frac{n-1}{\sum_{i=1}^n X_i}$$

$$T_3 = \sqrt{\frac{2n}{\sum_{i=1}^n X_i^2}}$$

$$T_4 = \left( \frac{6n}{\sum_{i=1}^n X_i^3} \right)^{\frac{1}{3}}$$

$$T_5 = \frac{\log(\frac{4}{3})}{X_{((\frac{n}{4}])}}$$

$$T_6 = \frac{\log(2)}{X_{((\frac{n}{2}])}}$$

$$T_7 = \frac{\log(3)}{X_{((\frac{2n}{3}])}}$$

$$T_8 = \frac{\log(4)}{X_{((\frac{3n}{4}])}}$$

où $(x]$ désigne le plus petit entier supérieur ou égal à $x$. (fonction *ceiling* de R).

1. Choisir une valeur de $\lambda$ et simuler 1000 échantillons de taille 100 de la loi exponentielle de paramètre $\lambda$. Calculer pour chacun de ces échantillons la valeur prise par les 8 estimateurs. Pour cela, on pourra utiliser la fonction suivante :

   ```
   > source("EstimParamModeleExponentiel.r")
   ```

   De plus, elle représente sur un même graphique les diagrammes en boîte (boxplot) des 8 estimateurs et elle superpose sur le même graphique la vraie valeur du paramètre, en rouge. Commenter la pertinence de chacun des estimateurs : lequel préfereriez-vous utiliser?

2. Calculer la moyenne empirique, et la variance empirique des 8 échantillons de taille 1000 ainsi obtenus. Utiliser la méthode de Monte Carlo pour en déduire une estimation du biais et de l'erreur quadratique de chacun des 8 estimateurs. On rappelle que pour un estimateur $T$, le biais est $b(\lambda) = E[T] - \lambda$ et l'erreur quadratique est $E\left[(T-\lambda)^2\right]$. La décomposition biais-variance stipule que $E\left[(T-\lambda)^2\right] = \mathrm{V}ar(T) + b(\lambda)^2$. Concernant $T_2$, en utilisant la densité de la loi Gamma (cf partie 2.1), on peut montrer que $E\left[(T_2-\lambda)^2\right] = \frac{\lambda^2}{n-2}$. Vérifier la cohérence de votre estimation de l'erreur quadratique de $T_2$.

3. Représenter des histogrammes pour les 8 estimateurs, avec les mêmes échelles. Représenter sur chaque graphique la valeur exacte de $\lambda$ par une ligne verticale rouge. Représenter par deux lignes verticales jaunes les bornes de l'intervalle de confiance de niveau 0.95 pour l'espérance de l'estimateur. Pour cela, on pourra utiliser la procédure suivante :

   ```
   m <- min(T)
   M <- max(T)

   bornes <- seq(from=m,to=M,le=30)
   za <- qnorm(0.975)
   var=apply(T,1,var)
   amp <- za*sqrt(var)/sqrt(n)
   moy=apply(T,1,mean)
   bi <- moy - amp
   bs <- moy + amp

   layout(matrix(1:8,2,4,byrow=TRUE));

   for (i in 1:2)
   {
   ```

```
  for (j in 1:4)
  {
    k = i+(j-1)*2

hist(T[,k],breaks=bornes,col="lightblue1",main =paste("Estimateur T",k,sep=""))

      abline(v=lambda,col="red")

      abline(v=bi[k],col="yellow")

      abline(v=bs[k],col="yellow")
  }
}
```

### 1.2 Une propriété de stabilité de la loi exponentielle

Dans cette partie, on souhaite vérifier numériquement le fait suivant. Si $X$ suit une loi exponentielle de paramètre $\lambda$ et $Y$ suit une loi exponentielle de paramètre $\mu$ avec $X$ et $Y$ indépendantes alors $Z = \min(X, Y)$ suit une loi exponentielle de paramètre $\lambda + \mu$. Cette observation est très utile pour déterminer l'espérance de vie d'un systme constitué de deux composants en série. Pourquoi ?

Pour cela, simuler un échantillon de taille 10000 de la exponentielle de paramètre $\lambda = 10$ puis un deuxième indépendant du premier de taille 10000 de la exponentielle de paramètre $\mu = 2$ (fonction `rexp` de `R`). Prenez le minimum des deux échantillon. Vérifier la propriété en traçant les deux graphiques suivant. Pour commencer, sur le même graphique, l'histogramme de vos points avec la densité de la loi exponentielle de paramètre $\lambda + \mu$. Ensuite, sur un autre graphique, la fonction de répartition empirique de vos points avec la fonction de répartition de la loi exponentielle de paramètre $\lambda + \mu$.

### 1.3 Convergence de la médiane

1. Générer un échantillon de taille $n = 100$ et de loi exponentielle de paramètre $\mu = 0.5$ (utiliser rexp()).

2. Calculer la médiane $M_n$ de l'échantillon (utiliser : median()). Calculer la médiane $m$ théorique de la loi exponentielle de paramètre $= 0.5$ (utiliser : qexp()).

3. Refaire $B = 150$ fois cette expérience et mettre les valeurs de $Z_n = \sqrt{n}\,(M_n - m)$ dans un vecteur de taille $B$. Estimer la variance $V_n$ de la médiane empirique en utilisant les $B$ expériences. Tracer lhistogramme des $Z_n$ et superposer la densité de la loi normale $N(0, 1)$. Que pouvez vous dire ?

## 2 Modèles de Gamma et de Weibull

Les lois exponentielles sont des cas particuliers de deux types de lois classiques plus généraux, couramment utilisés pour modéliser des durées de fonctionnement, les *lois Gamma* et les *lois de Weibull*.

## 2.1 Modèle de Gamma

La loi Gamma de paramètres $a > 0$ et $\lambda > 0$ a pour densité :

$$f(t) = \frac{\lambda^a}{\Gamma(a)} t^{a-1} e^{-\lambda t} \mathbb{1}_{t>0} \,,$$

où $\Gamma$ est la "fonction gamma", définie par : $\Gamma(a) = \int_0^{+\infty} t^{a-1} e^{-t} dt$.

La loi Gamma de paramètre 1 et $\lambda$ est la loi exponentielle de paramètre $\lambda$. Si $X$ suit la loi Gamma de paramètres $a$ et 1 alors $\frac{1}{\lambda} X$ suit la loi Gamma de paramètre $a$ et $\lambda$. De plus, la somme de $n$ variables aléatoires indépendantes de même loi exponentielle de paramètre $\lambda$ suit la loi Gamma de paramètre $n$ et $\lambda$. Cette loi porte alors le nom de loi d'Erlang.

Étudions le problème de l'estimation des paramètres $a$ et $\lambda$ à partir d'un $n$-échantillon de la loi Gamma de paramètres $a$ et $\lambda$. Soit $(X_1, ..., X_n)$ un échantillon de la loi Gamma de paramètres $a$ et $\lambda$. Considérons les estimateurs suivant de $a$ et $\lambda$ :

$$\widehat{a} = \frac{\overline{X}^2}{S^2}$$

$$\widehat{\lambda} = \frac{\overline{X}}{S^2} \,,$$

où $\overline{X}$ et $S^2$ sont respectivement la moyenne et la variance empirique de l'échantillon. Reprendre la partie 1.1 pour ce modèle Gamma afin d'étudier la qualité de ces estimateurs. On pourra utiliser la fonction suivante :

```
> source("EstimParamModeleGamma.r")
```

Elle calcule pour chacun des échantillons la valeur prise par ces estimateurs. De plus, elle représente sur un même graphique les diagrammes en boîte (boxplot) de ces estimateurs et elle superpose sur le même graphique la vraie valeur du paramètre, en rouge.

On completera cette etude par celle du risque quadratique, du biais et de la variance de ces estimateurs en s'inspirant de la partie 1.1, question 2.

Que pensez-vous de la qualite de ces estimateurs ? Quelle est l'influence de $n$ ?

## 2.2 Modèle de Weibull

La loi de Weibull de paramètres $a > 0$ et $\lambda > 0$, notée $W(a, \lambda)$, a pour densité

$$f(t) = \lambda a t^{a-1} e^{-\lambda t^a}$$

La loi $W(1, \lambda)$ est la loi exponentielle de paramètre $\lambda$. De plus, si $X$ suit la loi $W(a, 1)$ alors $\frac{1}{\lambda^{\frac{1}{a}}} X$ suit la loi $W(a, \lambda)$.

### 2.2.1 Simulation de la loi de Weibull

Le fait suivant peut servir de définition de la loi de Weibull. Si $X$ suit une loi exponentielle de paramètre 1 alors $W = \frac{X^{\frac{1}{a}}}{\lambda}$ suit une loi de Weibull $W(a, \lambda^a)$. Illustrer numériquement ce fait et en déduire un algorithme de simulation de la loi de Weibull. On pourra s'inspirer de la partie 1.2 et utiliser le code suivant :

```
> source("SimulationLoiWeibull.r")
```

Faire varier le paramètre de forme $a$ de cette procédure. Qu'observez-vous ? Utiliser le fait que la médiane d'une loi exponentielle est plus petite que 1 pour expliquer le fait que la loi se concentre en 0 pour de petites valeurs du paramètre de forme.

### 2.2.2 Estimation dans le modèle de Weibull

Soit $(X_1, ..., X_n)$ un échantillon de la loi de Weibull de paramètres $a$ et $\lambda$. Voici une méthode simple à mettre en oeuvre pour estimer les paramètres $a$ et $\lambda$ dans un tel modèle. Elle repose sur la régression linéaire simple au sens des moindres carrés. L'idée est la suivante. Pour $i = 1, \cdots, n$, la statistique d'ordre $X_{(i)}$ doit être proche (pour $n$ grand) du quantile d'ordre $\frac{i}{n}$ de la loi de Weibull de paramètres $a$ et $\lambda$. Or, ce quantile est calculable explicitement. On a donc

$$X_{(i)} \approx \left(-\frac{1}{\lambda}\log\left(1-\frac{i}{n}\right)\right)^{\frac{1}{a}}$$

soit

$$\log(X_{(i)}) \approx \frac{1}{a}\log\left(-\log\left(1-\frac{i}{n}\right)\right) + \frac{1}{a}\log\left(\frac{1}{\lambda}\right)$$

Posons $x_i = \log\left(-\log\left(1-\frac{i}{n}\right)\right)$ et $Y_i = \log(X_{(i)})$. Les points $(x_i, Y_i)$ devraient être proches de la droite d'équation $y = (1/a)x + (1/a)\log\left(\frac{1}{\lambda}\right)$. Les estimateurs des moindres carrés $A$ et $B$ pour la régression linéaire simple des $Y_i$ sur les $x_i$ sont des estimateurs de $1/a$ et $(1/a)\log\left(\frac{1}{\lambda}\right)$ respectivement. Donc, on propose les estimateurs suivants :

$$\widehat{a} = \frac{1}{A}$$

et

$$\widehat{\lambda} = e^{-\frac{B}{A}}\,.$$

Reprendre l'étude faite dans la partie 1.1 pour évaluer la qualité de ce procédé d'estimation. Pour l'obtention des boxplots des estimations, on pourra utiliser la procédure :

```
> source("EstimParamModeleWeibull.r")
```

Verifier graphiquement que, lorsque $n$ augmente (prendre successivement $n = 10, 100, 1000, 10000$), l'ajustement par une droite est de plus en plus pertinent. On peut alors remarquer que la qualité des estimateurs s'améliore nettement. On completera cette etude par celle du risque quadratique, du biais et de la variance de ces estimaieurs en s'inspirant de la partie 1.1, question 2.

Que pensez-vous de la qualite de ces estimateurs ?

### 2.2.3 Modélisation

La loi de Weibull permet de modéliser des durées de fonctionnement de systèmes dans lesquels la défaillance d'un seul composant suffit à provoquer la panne de tout système. Cette assertion repose sur le théorème suivant :

**Théorème 2.1** *Soient $X_1, \cdots, X_n$ $n$ durées de fonctionnement indépendantes et de même densité $f$. On suppose qu'il existe deux réels strictement positifs $a$ et $\lambda$ tels que, lorsque $t$ décroit vers $0$ :*

$$\frac{f(t)}{\lambda a t^{a-1}} \to 1\,.$$

*Notons $M_n$ la variable aléatoire :*

$$M_n = n^{\frac{1}{a}} \min(X_1, \cdots, X_n)\,.$$

*Alors la suite $(M_n)$ converge en loi lorsque $n$ tend vers $+\infty$ vers la loi de Weibull $W(a, \lambda)$*

Ce théorème indique que si la densité des $X_i$ est équivalente en $t = 0$ à la densité de la loi de Weibull $W(a, \lambda)$ alors le minimum des $X_i$ suit approximativement une loi de Weibull. La preuve est simple et repose sur la convergence des fonctions de répartition.

Pour comprendre ce résultat, pensons par exemple à un câble sous-marin. Coupons-le par la pensée en $n$ tronçons de longueur 1 km. Evidemment, la liaison est coupée dès qu'un des tronçons cesse de transmettre. Si $X_1, \cdots, X_n$ sont les durées de fonctionnement des $n$ tronçons, alors $\min(X_1, \cdots, X_n)$ est la durée de fonctionnement du câble. D'après le théorème précédent, elle suit approximativement une loi de Weibull.

Illustrer ce fait en supposant que les durées de fonctionnement de chaque tronçons suit une loi exponentielle de paramètre $\lambda$. Dans ce cas, l'hypothèse du théorème est vérifiée avec $a = 1$. Pour cela, on pourra fixer $n = 10000$ et calculer 1000 valeurs de $M_{10000}$. Vérifier la propriété en traçant les deux graphiques suivant. Pour commencer, sur le même graphique, l'histogramme de vos points avec la densité de la loi de Weibull $W(1, \lambda)$ . Ensuite, sur un autre graphique, la fonction de répartition empirique de vos points avec la fonction de répartition de la loi de Weibull $W(1, \lambda)$. Faire varier $n$ : le résultat est-il vrai uniquement pour les $n$ grands ? Pourquoi ?

# 3 Estimation dans le modèle uniforme : à vous de jouer.

Soit $(X_1, ..., X_n)$ un échantillon de la loi uniforme sur $[0, \theta]$ où $\theta$ est un paramètre inconnu. On note $(X_{(1)}, ..., X_{(n)})$ les statistiques d'ordre de l'échantillon. Elles vérifient :

$$X_{(1)} \leq ... \leq X_{(n)}\,.$$

On considère les estimateurs convergents (on dit aussi consistants) suivants du paramètre $\theta$.

$$T_1 = \frac{2}{n} \times (X_1 + ... + X_n)\ .$$

$$T_2 = \sqrt{\frac{3}{n} \times (X_1^2 + ... + X_n^2)}\,.$$

$$T_3 = \left(\frac{4}{n} \times \left(X_1^3 + ... + X_n^3\right)\right)^{\frac{1}{3}} .$$

$$T_4 = \left(\frac{3}{2n} \times \left(\sqrt{X_1} + ... + \sqrt{X_n}\right)\right)^{2} .$$

$$T_5 = \left(\frac{1}{2n} \times \left(\frac{1}{\sqrt{X_1}} + ... + \frac{1}{\sqrt{X_n}}\right)\right)^{-2} .$$

$$T_6 = \exp(1) \times (X_1 \times ... \times X_n)^{\frac{1}{n}} .$$

$$T_7 = 4 \times X_{(\left(\frac{n}{4}\right])} .$$

$$T_8 = 2 \times X_{(\left(\frac{n}{2}\right])} .$$

$$T_9 = \frac{3}{2} \times X_{(\left(\frac{2n}{3}\right])} .$$

$$T_{10} = \frac{4}{3} X_{(\left(\frac{3n}{4}\right])} .$$

$$T_{11} = \max\{X_1, ..., X_n\} .$$

$$T_{12} = \frac{n+1}{n} \max\{X_1, ..., X_n\} .$$

Reprendre les questions de la partie 1.1 pour ce modèle et ces estimateurs de $\theta$. Pour l'estimateur $T_6$, on pourra utiliser :

```
T6 <- exp(1)*(apply(X,1,prod)^(1/n))
```

Commenter la pertinence de chacun des estimateurs : lequel préfereriez-vous utiliser?