M1 MAPI3 2022-2023
Statistiques

Université Paul Sabatier
philippe.berthet@math.univ-toulouse.fr
adrien.mazoyer@math.univ-toulouse.fr
maxime.el_masri@math.univ-toulouse.fr

# TP 1
# Estimation – vraisemblance – illustration par simulation

**Remarque** Ce TP est dédié à l'introduction du langage R à travers l'illustration numérique de propriétés d'estimateurs.
Si le langage R est nouveau pour vous, un "Guide" est accessible sur Moodle. Celui-ci a été rédigé initialement pour des cursus d'Économie, dépourvu d'informatique, et est donc très accessible.
Coder en R fonctionne globalement comme en Python, mais les fonctions classiques, les classes d'objet, etc sont légèrement différents. Le Guide est à justement pour cela : il n'est pas du tout nécessaire de le parcourir en entier, en revanche la majeure partie des informations dont vous aurez besoin dans les TPs seront soit indiquées dans les sujets, soit accessibles dans ce Guide. Servez-vous de la table des matières et/ou de la fonctionnalité de Recherche de votre lecteur pdf pour trouver l'information.
Également, n'hésitez pas à vous servir d'internet qui regorge de réponses à la majorité des questions que vous vous poserez ! Vous pouvez par exemple consulter les pages suivantes (liste non-exhaustive)

- `https://www.statmethods.net/index.html`
- `http://www.sthda.com/english/`
- `https://r4ds.had.co.nz/index.html`

ainsi que les sites classiques type `https://stackoverflow.com/` etc.

## 1 Fiabilité d'un circuit électrique (Ex7 du TD1)

On rappelle le contexte : on dispose d'un circuit électrique constitué de deux diodes montées en série. Chaque diode a une durée de vie, modélisée par une loi exponentielle de paramètres respectifs $\theta_1$ et $\theta_2$, et on cherche à estimer ces 2 paramètres au vu deux types d'échantillons.

### 1.1 Données "incomplètes"

On se place d'abord dans le premier cas, où ne peut accéder qu'aux durées de vie du circuit global sans connaître la diode causant la défaillance. Lorsqu'on se place dans ce cas, on dispose donc de données qu'on supposera issues d'une loi exponentielle de paramètre $g(\theta) = \theta_1 + \theta_2$. Un échantillon de la loi exponentielle s'obtient simplement grâce à la fonction `rexp`, tandis que sa densité se calcule avec `dexp`.

```
## Simulation d'un échantillon de taille 10 selon la loi exponentielle de taux 1
rexp(10, rate = 1)
```

```
##  [1] 0.06175786 0.49630969 0.32826760 1.56525201 0.28118093 2.10307837
##  [7] 1.50792060 0.49008045 0.40373731 0.19876341
```

```
## Calcul de la densité en 0.2 pour la loi exponentielle de taux 1
dexp(0.2, rate = 1)
```

```
## [1] 0.8187308
```

1. Rappelez pourquoi l'estimation de $\theta_1$ et $\theta_2$ n'est pas possible dans ce contexte.

On a montré en TD que l'EMV de $g(\theta)$, basé sur un échantillon $(Z_1, \dots, Z_n)$, est donné par

$$\widehat{g(\theta)}_n = \frac{1}{\overline{Z}}$$

Mais on va implémenter nous même le calcul de l'EMV. L'optimisation d'une fonction dans R peut se faire via la fonction `optim`, qui permet de chercher le **minimum** d'une fonction. Par exemple, pour chercher le minimum de la fonction $f(x) = x^2$, il suffit d'exécuter le code ci-dessous

```
f <- function(x) {
        return(x^2)
}
## L'argument 'par' de la fnction optim permet d'initialiser le point où est évalué f avant de chercher le mini
optim(par = 0.1, fn = f)
```

```
## $par
## [1] -8.326673e-17
##
## $value
## [1] 6.933348e-33
##
## $counts
## function gradient
##       32       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

Parmi toutes les informations renvoyées par la fonction `optim`, seul l'attribut `par` nous intéresse :

```
optim(par = 0.1, fn = f)$par
```

```
## [1] -8.326673e-17
```

Il est possible de restreindre la recherche du minimum à un intervalle donné (il faut préciser pour cela l'argument `method` comme ci-dessous)

```
optim(par = 0.1, fn = f, lower = 2, upper = 3, method = "L-BFGS-B")$par
```

```
## [1] 2
```

2. À l'aide de la fonction `optim`, implémentez une fonction `EMV_theta` prenant en entrée un échantillon et renvoyant l'EMV de $g(\theta)$ associé. Vérifiez sur un échantillon que vous obtenez bien la valeur recherchée !

Pour finir, on va vérifier le bon comportement de notre estimateur. Les propriétés théoriques qui nous intéressent sont celles associés à l'Exercice 1 en annexe.

3. Pour illustrer numériquement ces propriétés, on va effectuer une étude de simulation comme décrit ci-dessous
   (i) Se fixer une valeur pour $g(\theta)$, et une taille d'échantillon $n$.
   (ii) Simuler un échantillon de taille $n$ de dudrée de vie.
   (iii) Estimer $g(\theta)$ avec la fonction `EMV_theta`.
   (iv) Répéter (i) à (iii) un nombre $N$ de fois, pour obtenir un vecteur de taille $N$ de différentes réalisations de l'EMV de $g(\theta)$.
4. Vérifier alors les propriétés sur le biais de l'estimateur, en faisant varier $g(\theta)$ et $n$.
5. Illustrez graphiquement le résultat théorique sur la loi de notre estimateur.

## 1.2 Données "complètes"

On se place à présent dans le cas où les données considérées renseignent la durée de vie du circuit ainsi que la diode responsable de la défaillance, càd la question **d.** de l'Exercice 7 du TD1.

1. Créez une fonction `rcircuit` renvoyant cette fois ce type de données, en prenant en entrée
   - `n` : la taille de l'échantillon
   - `theta1` : le paramètre de durée de vie pour la diode 1
   - `theta2` : le paramètre de durée de vie pour la diode 2

   Cette fonction renverra donc une matrice à `n` lignes et 2 colonnes, la première contenant les réalisations de la variable $Y$, la 2ème celles de la variable $Z$.
2. Implémente dans une fonction `dcircuit` le calcul de la densité jointe du couple $(Y, Z)$, où $Z$ correspond à la durée de vie du circuit, et $Y$ indique la cause de la défaillance : où $Y = 1$ si c'est la diode 1, $Y = 0$ sinon.
3. Implémentez dans une fonction `EMV_theta_v2` le calcul de l'EMV du vecteur $(\theta_1, \theta_2)$. Assurez-vous de bien retrouver le résultat théorique établi en TD.
4. Rappelez pourquoi cet estimateur est asymptotiquement sans biais ? Illustrez numériquement cette propriété en prenant exemple sur la partie précédente.

# 2 Loi de Pareto

La loi de Pareto est accessible dans R, en chargeant le paquet `EnvStats` (voir Section 13 du Guide pour l'installation de paquets), sur le même modèle que les autres lois classiques (voir `?rpareto`).

On va comparer les performances des 3 estimateurs construits dans l'Exercice 2.

1. Commencez par implémenter trois fonctions distinctes, une pour chaque méthode d'estimation, qui pour un échantillon vous renvoie l'estimation du paramètre $\theta$.
2. En suivant la méthode décrite dans le point 4 de la partie 1.1, illustrez numériquement les propriésts des trois estimateurs. Vous prendrez bien garde à
   - Utiliser les mêmes échantillons simulés pour les 3 estimateurs
   - Considérer les différents cas de valeur du paramètre $\theta$.
3. Comparez les trois estimateurs en terme de biais et variance.

# Annexe TP 1
# Exercices théoriques

## Exercice 1

Soit un échantillon issu de $n$ variables $X_1, \dots, X_n$ iid selon la loi exponentielle $\mathcal{E}(\lambda)$. Une propriété de la loi exponentielle est la suivante : si $X_1, \dots, X_n$ sont toutes iid selon une loi $\mathcal{E}(\lambda)$, alors $\sum_{i=1}^{n} X_i$ suit la loi $\Gamma(n, \lambda)$, dont la densité est donnée par

$$g_\lambda(x) = \lambda^n \frac{x^{n-1}}{(n-1)!} \exp(-\lambda x) \mathbb{1}[x > 0]$$

1. Quelle est la loi de $\overline{X}$ ?
2. Calculer l'estimateur par la méthode des moments $\widehat{\lambda}_n$ de $\lambda$.
3. Montrer que
$$\mathbb{E}_\lambda\left[\widehat{\lambda}_n\right] = \frac{n}{n-1}\lambda \quad \text{et} \quad \mathbb{V}_\lambda\left[\widehat{\lambda}_n\right] = \frac{(n\lambda)^2}{(n-1)^2(n-2)}$$
4. Est-il sans biais ? Asymptotiquement sans biais ? Convergent ?
5. Montrer qu'on a la convergence en loi suivante
$$\sqrt{n}\left(\frac{1}{\overline{X}} - \lambda\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \lambda^2)$$

## Exercice 2

On considére à présent les lois de Pareto, caractérisées par

$$\mathbb{P}_\theta[X \leq x] = \left(1 - \frac{1}{x^\theta}\right) \mathbb{1}[x \geq 1]$$

où $\theta > 0$.

1. Montrez que pour une variable $X$ distribuée selon une loi de Pareto de paramètre $\theta$, on a que
$$\begin{aligned} \mathbb{E}_\theta[X] &= \frac{\theta}{\theta - 1} \qquad \theta > 1 \\ \mathbb{V}_\theta[X] &= \frac{\theta}{(\theta-1)^2(\theta-2)} \qquad \theta > 2 \\ \mathrm{Me}_\theta[X] &= 2^{1/\theta} \end{aligned}$$
où $\mathrm{Me}_\theta$ représente la médiane.

On va s'intéresser au problème d'estimation de $\theta$ à partir d'un échantillon issu de $X_1, \dots, X_n$ iid selon une loi de Pareto de paramètre $\theta$. On va pouvoir identifier trois estimateurs distincts.

2. Le premier estimateur, qu'on notera $\hat{\theta}_{1,n}$ se calcule grâce à la méthode des moments.
   (a) Montrer qu'avec cette méthode on obtient
   $$\hat{\theta}_{1,n} = \frac{\overline{X}}{\overline{X} - 1}$$
   (b) Montrer que cet estimateur vérifie les propriétés suivantes
   $$\lim_{n\to\infty} \hat{\theta}_{1,n} = \theta \qquad \text{et} \qquad \sqrt{n}\left(\hat{\theta}_{1,n} - \theta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \frac{\theta(\theta-1)^2}{\theta - 2}\right)$$
3. Le deuxième estimateur, qu'on notera $\hat{\theta}_{2,n}$, s'exprime à partir de la médiane empirique $M_n = X_{(\lfloor n/2 \rfloor)}$ (avec $X_{(1)} \leq \cdots \leq X_{(n)}$ l'échantillon ordonné).

(a) Montrer que

$$\hat{\theta}_{2,n} = \frac{\log(2)}{\log(M_n)}$$

(b) On donne les propriétés de la médiane empirique

$$\lim_{n\to\infty} M_n = \mathrm{Me}(X) \qquad \text{et} \qquad \sqrt{n}\left(M_n - \mathrm{Me}_\theta(X)\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \frac{1}{4 f_\theta(\mathrm{Me}[X])^2}\right)$$

avec $f_\theta$ la densité de la loi de Pareto de paramètre $\theta$. Montrer alors que l'estimateur $\hat{\theta}_{2,n}$ vérifie les propriétés suivantes

$$\lim_{n\to\infty} \hat{\theta}_{2,n} = \theta \qquad \text{et} \qquad \sqrt{n}\left(\hat{\theta}_{2,n} - \theta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \frac{\theta^2}{\log(2)^2}\right)$$

4. Le 3ème estimateur de $\theta$ est l'EMV.

(a) Montrez que celui-ci est donné par

$$\hat{\theta}_{3,n} = \frac{n}{\sum_{i=1}^{n} \log(X_i)}$$

(b) Après avoir étudié la loi de $W = \log(X)$ lorsque $X$ suit une loi de Pareto de paramètre $\theta$ en déduire que

$$\mathbb{E}_\theta\left[\hat{\theta}_{3,n}\right] = \frac{n}{n-1}\theta\,, \qquad \mathbb{V}_\theta\left[\hat{\theta}_{3,n}\right] = \frac{(n\theta)^2}{(n-1)^2(n-2)}$$

et qu'on a la convergence en loi suivante

$$\sqrt{n}\left(\hat{\theta}_{3,n} - \theta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \theta^2)$$