TP 6 : Méthode des moments VS méthode du maximum de vraisemblance

Ce TP étudie l'estimation par la méthode des moments (que l'on a déjà traitée sans la nommer) et l'estimation par maximum de vraisemblance, en montrant que ces deux méthodes peuvent conduire à des estimateurs différents. On applique ensuite ces estimateurs sur des données réelles.

Estimation d'un paramètre d'une power law.

La *power law* de paramètre $a > 1$ sur l'intervalle $[1, +\infty[$, notée $PL(a)$, est la loi de densité

$$f_a(x) = \frac{a-1}{x^a}\mathbf{1}_{x\geq 1}.$$

(on admet que c'est bien une densité)

Première partie : quelques simulations

1. Pour $a > 0$, calculer la fonction de répartition de la loi $PL(a)$. Ensuite, *un utilisant uniquement variables uniformes sur* $[0,1]$, simuler un échantillon de taille $n = 10^4$ de loi $PL(8.9)$. Représenter l'histogramme associé, et superposer la densité théorique de la loi $PL(8.9)$.

2. On donne le résultat plus général suivant : si $X \sim PL(a)$ et $Y \sim PL(b)$ avec X et Y indépendantes, alors $\min(X,Y) \sim PL(a+b-1)$. Illustrer ce phénomène pour $a = 3.4$ et $b = 6.2$. *Indication : pour calculer le minimum terme à terme de deux vecteurs, on pourra utiliser* `pmin`.

3. Pour $X \sim PL(a)$, quelle loi semble suivre $\log(X)$? Donner une illustration graphique de cette conjecture.

Deuxième partie : deux estimateurs Dans la suite, on suppose que l'on a un échantillon $X_1, \ldots, X_n$ de loi $PL(a)$.

4. Proposer un estimateur de a, noté $\hat{a}_{MM}$, utilisant la moyenne emprique $\bar{X}_n$. *C'est l'estimateur de la méthode des moments d'ordre 1 (1 car il utilise le moment d'ordre 1).* Quelle hypothèse doit-on faire sur a pour que cet estimateur converge ?

5. Ecrire la vraisemblance $L(a, X_1, \ldots, X_n)$ pour tout $a > 0$. Passer au log et optimiser cette log-vraisemblance pour trouver l'estimateur $\hat{a}_{MV}$ du maximum de vraisemblance.

6. Montrer que $\hat{a}_{MV}$ est convergent pour tout $a > 1$. On pourra admettre que la conjecture faite à la question 3 est vraie.

7. On cherche à représenter la normalité asymptotique des deux estimateurs. Pour $a = 3.9$, simuler $N = 5000$ échantillons de taille $n = 10^4$, puis tracer *sur le même graphique* les histogrammes de $\sqrt{n}(\hat{a}_{MM} - a)$ et de $\sqrt{n}(\hat{a}_{MV} - a)$. *Indication : on utilisera* `add=TRUE` *pour le deuxième histogramme. De plus, pour donner de la transparence aux couleurs afin de mieux superposer, on pourra utiliser en paramètre* `col = rgb(red = 1, green = 0, blue = 0, alpha = 0.5)`. *Plus* `alpha` *est proche de* 0*, plus le graphique est transparent.*

Quel semble être le meilleur estimateur, au vu de cette loi limite ? Justifier.

8. Confirmer la réponse faite à la question précédente en affichant les risques estimés de ces estimateurs.

9. Recopier et relancer le code de la question 7, cette fois-ci pour $a = 1.8$. Que se passe-t-il, et pourquoi ?

10. Au vu de ces résultats, donnez deux arguments (scientifiques) qui motivent le choix d'un des deux estimateurs.

Troisième partie : application à des données réelles Pour cette partie, vous téléchargerez le fichier 'moby_dick.csv' à la page suivante https://lganassali.github.io/cours_TP_stats.html, et vous le placerez dans le même dossier que votre TP6. Ainsi, il sera très simple de lire les données, avec le code suivant :

```r
data = read.csv("moby_dick.csv")
head(data)
```

```
##   occurrences
## 1       14086
## 2        6414
## 3        6260
## 4        4573
## 5        4484
## 6        4040
```

Ce jeu de données recense le nombre d'occurences de chaque mot présent dans le roman Moby Dick d'Herman Melville (roman de 1851, grand classique de la littérature américaine). Il y a en tout 18855 mots différents dans ce roman de près de 600 pages. On a enlevé le label des mots, on garde juste le nombre d'occurences.

11. A quel mot correspond la première donnée, à votre avis ?

12. On extrait notre vecteur de données X comme suit :

```r
X = data$occurrences
```

On souhaite montrer que X suit approximativement une loi $PL(a)$. On déterminera une valeur approchée de ce paramètre a avec l'estimateur du maximum de vraisemblance, et on superposera les courbes des fonctions de répartition.

13. Le graphique ci-dessus est-il satisfaisant ? Pourquoi ? Une autre façon de représenter la proximité en loi est d'utiliser les résultats de la question 3 : réaliser le plot de la fonction de répartition empirique du log-échantillon. Superposer la courbe adéquate. Commenter.

14. Relancer le code de la question 13 en estimant a avec l'estimateur des moments. Le fit est-il meilleur ?

Comme conclusion, on pourra retenir la phrase suivante :

"In theory, there is no difference between theory and practice. In practice, there is."