Master 2 Data Science — Année universitaire 2022-2023
Gilles Stupfler — Email : gilles.stupfler@univ-angers.fr

---

**Illustrations : Lois à queue lourde**

Les jeux de données pertinents pour cette session d'illustrations sont :

- Données de pluie d'Orlando, Floride, USA, voir `Florida_rainfall.RData`,
- Des données financières de l'indice CAC40, voir `CAC40.csv`,
- La base de données Group Medical Insurance Large Claims de la Society of Actuaries (SOA), voir le package R `ReIns`,
- Des données du recensement américain de 2010 classant les villes américaines par population, voir `census_USA_2010.xlsx`.

Dans chaque cas, il est intéressant de coder soi-même les méthodes et de se comparer aux procédures classiques des packages `evir`, `evd`, `evt0` et `extRemes`. Vous pourrez d'abord :

1. Implémenter les estimateurs de Hill et Weissman, avec une sortie sous forme de vecteurs.
2. Implémenter une méthode de diagnostic (QQ-plot...)
3. Télécharger quelques packages et les comparer avec vos implémentations.

### 1. Données pluviométriques

La variable d'intérêt est `sum_rain_2m_inches`, la quantité totale de pluie journalière enregistrée par une station météo à Orlando, Floride, USA, pendant la période annuelle de tempêtes qui dure d'août à octobre.

1. Représenter les données en ignorant la structure temporelle puis en tenant compte de cette structure. Y a-t-il de la saisonnalité, de l'autocorrélation, de la non-stationnarité ?
2. Peut-on proposer un modèle statistique raisonnable pour la totalité de la distribution des données ?
3. Montrer une preuve de la présence d'une queue lourde dans les données.

4. Estimer l'indice des valeurs extrêmes des données en justifiant le choix du/des paramètres utilisés.

5. Calculer un estimateur de quantile extrême de la variable d'intérêt aux niveaux 0.99, 0.995 et 0.999. Qu'en pensez-vous ?

6. Peut-on calculer un intervalle de confiance pour l'indice des valeurs extrêmes et les quantiles extrêmes ? Si oui, pourquoi et comment ? Sinon, peut-on suggérer une méthode qui le permettrait ?

**2. Données CAC40**

Les données financières ne sont généralement pas stationnaires (inflation...) On isolera d'abord la variable `Close`, désignant le prix de clôture de l'indice, puis on construira le log-retour journalier $X_t$ : si $S_t$ est le prix de clôture au jour $t$, le log-retour journalier au jour $t$ est $\log(S_t/S_{t-1})$. C'est cette variable $X_t$ qui sera la variable d'intérêt.

1. Représenter les données en ignorant la structure temporelle puis en tenant compte de cette structure. Y a-t-il de la saisonnalité, de l'autocorrélation, de la non-stationnarité ?

2. Peut-on proposer un modèle statistique raisonnable pour la totalité de la distribution des données ?

3. Montrer une preuve de la présence d'une queue lourde dans les données.

4. Estimer l'indice des valeurs extrêmes des données en justifiant le choix du/des paramètres utilisés.

5. Calculer un estimateur de quantile extrême de la variable d'intérêt aux niveaux 0.995 et $1 - 1/n$. Qu'en pensez-vous ?

6. Peut-on calculer un intervalle de confiance pour l'indice des valeurs extrêmes et les quantiles extrêmes ? Si oui, pourquoi et comment ? Sinon, peut-on suggérer une méthode qui le permettrait ?

**3. Données SOA**

La seule variable présente ici est le montant d'un remboursement dû à des frais médicaux dépassant 25 000 $ en 1991 aux USA. On obtient ces données en écrivant `data(soa)` après avoir chargé le package `ReIns`.

1. Représenter les données. Peut-on proposer un modèle statistique raisonnable pour la totalité de la distribution des données ?

2. Montrer une preuve de la présence d'une queue lourde dans les données.

3. Estimer l'indice des valeurs extrêmes des données en justifiant le choix du/des paramètres utilisés.

4. Calculer un estimateur de quantile extrême de la variable d'intérêt aux niveaux 0.995 et $1-1/n$. Qu'en pensez-vous ? Le niveau 0.995 peut-il être considéré comme extrême ici ?

5. Peut-on calculer un intervalle de confiance pour l'indice des valeurs extrêmes et les quantiles extrêmes ? Si oui, pourquoi et comment ? Sinon, peut-on suggérer une méthode qui le permettrait ?

**4. Données de recensement US**

La variable d'intérêt est la taille de la population des villes américaines de plus de 50 000 habitants en 2010.

1. Représenter les données. Peut-on proposer un modèle statistique raisonnable pour la totalité de la distribution des données ?

2. Montrer une preuve de la présence d'une queue lourde dans les données.

3. Estimer l'indice des valeurs extrêmes des données en justifiant le choix du/des paramètres utilisés.

4. Calculer un estimateur de quantile extrême de la variable d'intérêt aux niveaux 0.99, 0.995 et 0.999. Qu'en pensez-vous ? Pouvez-vous les interpréter ?

5. Peut-on calculer un intervalle de confiance pour l'indice des valeurs extrêmes et les quantiles extrêmes ? Si oui, pourquoi et comment ? Sinon, peut-on suggérer une méthode qui le permettrait ?