Statistique descriptive avec R
TP 6

Nous allons travailler sur des données d'assurance du paquet `CASdatasets`.

1. Le paquet est trop volumineux pour être hébergé sur le CRAN en raison de données lourdes (236 Mo). Commencez par l'installer à l'aide du code suivant, puis vérifiez qu'il se charge correctement.

```
install.packages("CASdatasets", repos = "https://cas.uqam.ca/pub/",
    type="source")
library(CASdatasets)
```

2. Chargez les jeux de données `freMTPLfreq` et `freMTPLsev` avec la fonction `data`, puis convertissez-les en objets `data.table` sans effectuer de copie. Le premier jeu de données décrit les polices d'assurance automobile, avec leurs caractéristiques et le nombre de sinistres associés ; le second contient les détails des sinistres individuels correspondants.

3. Commençons par le jeu de données `freMTPLfreq`. La colonne `ClaimNb` indique le nombre de sinistres observés par police. Affichez le tableau des fréquences pour chaque nombre de sinistres, puis calculez la moyenne et la variance de cette variable.

4. La colonne `Exposure` correspond à la durée d'observation de chaque assuré (en années). Pour un assuré i d'exposition E_i, on s'attend à λ sinistres par unité d'exposition, soit $\mathbb{E}(N_i) = \lambda E_i$ où N_i est le nombre de sinistres. Sous une loi de Poisson $N_i \sim \mathcal{P}(\lambda E_i)$, on aurait aussi $Var(N_i) = \lambda E_i$.

Proposez deux estimateurs de λ : $\widehat{\lambda}_n^1$ basé sur la moyenne empirique (moyenne pondérée) et $\widehat{\lambda}_n^2$ basé sur la variance (sous hypothèse poissonienne). Concluez, sans test formel, si les données semblent suivre a priori une loi de Poisson.

5. À l'aide de $\widehat{\lambda}_n^1$, calculez les fréquences théoriques du nombre de sinistres $N_i \in \{0, 1, 2, 3, 4\}$ sous l'hypothèse $N_i \sim \mathcal{P}(\widehat{\lambda}_n^1 E_i)$. Stockez-les dans un vecteur nommé `freq_poisson`. Récupérez les fréquences empiriques (calculées à la question 3) dans `freq_empirique`.

6. Représentez graphiquement `freq_empirique` et `freq_poisson` pour les comparer. Commentez les écarts observés.

7. En assurance, on suppose souvent l'indépendance entre la fréquence (N_i) et la gravité moyenne des sinistres ($\frac{S_i}{N_i}$ pour $N_i > 0$). Vérifiez empiriquement cette hypothèse en calculant la corrélation entre ces deux quantités.

Pour cela :

- Agrégez `freMTPLsev` par `PolicyID` pour obtenir $S_i =$ `sum(ClaimAmount)` pour tout i tel que $N_i > 0$.

- Effectuez une jointure à gauche avec `freMTPLfreq` sur la clé `PolicyID` afin d'ajouter `ClaimNb`.
- Calculez $\frac{S_i}{N_i}$ pour $N_i > 0$, puis la corrélation avec N_i.

La syntaxe de jointure dans `data.table` est : `X[Y, on = "clé", col_dest := col_source ]`.

8. Passons maintenant à l'analyse des coûts des sinistres (variable `ClaimAmount` dans `freMTPLsev`). Tracez l'histogramme de leur répartition avec 100 barres (`breaks = 100`) :

— pour l'ensemble des données ;

— en tronquant au quantile 99,5 % (pour atténuer l'effet des valeurs extrêmes) ;

— en tronquant au quantile 95 %.

Commentez les particularités observées (ex. : pics anormaux).

9. Pour analyser la gravité des sinistres en fonction des variables explicatives, ajoutez celles de `freMTPLfreq` à `freMTPLsev` via une jointure à gauche sur `PolicyID`. Attention : convertissez d'abord `PolicyID` en entier dans `freMTPLfreq` pour éviter les erreurs de type.

10. Calculez la matrice de corrélations entre *toutes* les variables (sauf `PolicyID`) :