Statistique descriptive avec R
TP 5

Nous allons travailler sur les données américaines `flights14` du Bureau des statistiques des transports (*Bureau of Transportation Statistics*) pour les vols du départ de New York de janvier à octobre 2014.

1. Importez avec le paquet **`data.table`** et explorez vos données.

2. Il ne faut jamais parcourir les lignes une à une dans un **`data.table`**, et de même dans un **`data.frame`**. La fonction **`system.time`** permet de mesurer le temps de calcul effectif d'un bloc de code. Exécutez les deux codes suivants. Testez également avec l'objet en **`data.frame`** (ne pas oublier de vous remettre en **`data.table`** pour la suite). Comparez à une solution efficace que vous proposerez.

```
# Code 1 avec extraction de ligne puis sélection
system.time({
    s = 0
    for(i in 1:nrow(flights))
        if(flights[i, ]$origin == "JFK")
            s = s + flights[i, ]$dep_delay
})

# Code 2 avec extraction de ligne et sélection
system.time({
    s = 0
    for(i in 1:nrow(flights))
        if(flights[i, "origin"] == "JFK")
            s = s + flights[i, "dep_delay"]
})
```

3. Récupérez la moyenne et l'écart-type du retard au départ en utilisant la fonction **`tapply`**.

4. Faites la question précédente en utilisant la syntaxe de **`data.table`**. On utilisera la syntaxe `mon.data.table[, f(colonneB), by=colonneA]` où on agrège par `f` la `colonneB` selon les modalités de la `colonneA`. Pour obtenir ensemble la moyenne et l'écart-type (mais aussi pour nommer une colonne au renvoi), on utilisera la syntaxe plus générale `mon.data.table[, .( nom1 = f(colonneB), nom2 = g(colonneC)), by=colonneA]`.

5. On rappelle que si les $(X_i)_{1\leq i\leq n}$ sont des variables aléatoires i.i.d. et de carré intégrable, nous avons le théorème central limite sur $\overline{X}_n$ la moyenne empirique avec $\widehat{\sigma}_n$ l'écart-type empirique :

$$\sqrt{n}\frac{\overline{X}_n}{\widehat{\sigma}_n} \xrightarrow[n\to+\infty]{loi} \mathcal{N}(0,1),$$

ce qui permet de construire un intervalle de confiance asymptotique :

$$\mathbb{P}\left(\mathbb{E}(X) \in \left[\overline{X}_n \pm q_{1-\frac{\alpha}{2}} \frac{\widehat{\sigma}_n}{\sqrt{n}}\right]\right) \underset{n\to+\infty}{\longrightarrow} 1-\alpha,$$

où $q_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ d'une loi normale centrée réduite. Pour chaque aéroport de départ donnez la moyenne et les deux bornes de l'intervalle de confiance au niveau de 95% en exploitant la syntaxe de `data.table` et commentez.

6. Construisez un vecteur de couleurs où à chaque aéroport de départ on associe une couleur.

7. Affichez `arr_delay` en fonction de `dep_delay` en mettant une couleur différente en fonction de l'aéroport de départ. Ajoutez une légende. Si vous enregistez le graphique, en quel format est-ce préférable ?

8. Calculez la corrélation linéaire entre `dep_delay` et `arr_delay` en fonction de l'aéroport d'origine. Puis calculez la corrélation entre `dep_delay` et `arr_delay` − `dep_delay` (qui est le retard ou l'avance additionnelle après départ) en fonction de l'origine et pour les données complètes. Que remarquez-vous ?

9. Affichez le graphique précédent, mais en comparant cette fois-ci à `arr_delay` − `dep_delay` à `dep_delay`.

10. Ajoutez une colonne `dep_delay_group` qui vaut `"<=60"` si le retard au départ est inférieur à 60, `">60&<=300"` s'il est entre 60 et 180 et `">300"` s'il est supérieur à 180.

11. Pour `arr_delay` − `dep_delay`, calculez la moyenne, l'écart-type, la corrélation linéaire en fonction de `arr_delay` − `dep_delay`, et le nombre de données pour chaque `dep_delay_group_name` .

12. Affichez `arr_delay` − `dep_delay` en fonction de `dep_delay` avec une couleur différente par `dep_delay_group` et commentez.