


## TP2 : Statistique exploratoire

Ce TP a pour objectif de mener l'étude descriptive uni et bi-dimensionnelle du jeu de données **eauxTP** disponible sous Moodle. Vous rédigerez les réponses et vos observations dans un rapport (au format pdf ou html) grâce à R Markdown. N'oubliez pas d'interpréter les résultats !!!

### 1 Données

Un client vous demande de faire une analyse descriptive du jeu de données **eauxTP.csv** disponible sur la page moodle de l'UF. Ce jeu de données comprend la description de différentes marques d'eaux commercialisées en France. Elles sont décrites par les variables ci-dessous :

- Nom : le nom complet de l'eau inscrit sur l'étiquette.
- Acro : Le nom en abrégé.
- Pays : le pays d'origine identifié par les lettres de l'immatriculation automobile officielle.
- Type : 0 pour une eau de source et 1 pour une eau minérale.
- Nature : 0 pour une eau plate et 1 pour une eau gazeuse.
- leur teneur (en mg/litre) en ions calcium (Cal), magnésium (Mag), sodium (Sod), potassium (Pot), sulfates (Sul), nitrates (Nit), carbonates (Car) et chlorures (Chl).

Pour une prise en charge des données sous , utilisez la commande


```
> eaux <- read.table("eauxTP.csv", header=TRUE)
> head(eaux)
```

- i. Quelle est la dimension des données, i.e combien a-t-on d'individus et de variables ? Vous pouvez vous aider de la commande `dim()`.
- ii. `eaux` est-il de type `data.frame` (commande `is.data.frame()`) ? En plus du tableau de données, il possède certains attributs que l'on peut voir à l'aide de `attributes(eaux)` et `str(eaux)`. Quels sont-ils ?

Chaque attribut peut aussi être retrouvé individuellement. Tapez par exemple

```
> attributes(eaux)
> attributes(eaux)$names
```

- iii. Le nom d'une variable ou d'un individu peut se substituer à un indice. Vous pouvez le vérifier en tapant les commandes
 

```
> eaux$Cal[1] ; eaux[1,"Cal"] ; eaux[1,6]
```
- iv. Quelle est la nature de chaque variable ? Faut-il modifier leur nature dans le logiciel  ?

### 2 Etude statistique unidimensionnelle

#### 2.1 Pour une variable qualitative

Nous nous intéressons maintenant aux trois variables qualitatives. Ces dernières **doivent** être considérées comme des facteurs à plusieurs modalités.

- i. Stockez la variable "Pays" dans l'objet `Pays` et vérifiez qu'il est bien sous forme de facteur. A l'aide des commandes `table` et `summary`, donnez un résumé de cette variable. Quels sont les modalités du facteur ?
- ii. Résumez graphiquement la variable "Pays" par une représentation par secteurs (`pie()`) ou un diagramme en barres (`barplot()`). Interprétez les résultats.

#### 2.2 Pour une variable quantitative

Stockez dans un vecteur que l'on appellera `Nitrate` les teneurs en Nitrate.

### 2.2.1 Indices statistiques

- i. Que calculent les commandes `mean()`, `median()`, `var()`, `sd()`, `range()` ? Calculez l'étendue des données.
- ii. Etudiez les sorties des commandes `summary(Nitrate)` et `quantile(Nitrate)`. Donnez également l'écart interquartile et les valeurs adjacentes.

### 2.2.2 Représentations graphiques

Le but de la statistique exploratoire est de synthétiser, résumer et structurer l'information contenue dans des données. On utilise pour cela des représentations de données sous forme de tableaux ou de graphiques.

- i. La commande `hist(var)` permet de représenter l'histogramme d'une variable `var` donnée. Tapez `H<-hist(Nitrate)` et commentez les différents attributs de `H`. Exploitez les options `freq` et `breaks` de `hist`.
- ii. Représentez un boxplot de la variable "Nitrate" via la commande `boxplot()`. Que remarquez-vous ? Comparez avec les valeurs adjacentes calculées à la section 2.2.1.
- iii. En vous aidant de l'aide de R et des résultats de la section précédente, expliquez ce que représentent les différentes parties du graphique. Tapez `B<-boxplot(Nitrate)` et commentez les différents attributs de `B`.

## 3 Analyse bidimensionnelle

### 3.1 Entre deux variables quantitatives

- i. Calculez la matrice de corrélation des variables quantitatives (commande `cor()`) ou la matrice de variance-covariance (commandes `var()` ou `cov()`). Représentez graphiquement les corrélations à l'aide de la fonction `corrplot()` de la librairie `corrplot`. Vous pourrez utiliser l'option `method="ellipse"` pour une meilleure lisibilité.
- ii. Interprétez les résultats.
- iii. Représentez graphiquement la teneur en calcium en fonction de la teneur en sulfates à l'aide de la commande `plot`. Au vu de ce graphique, les variables sont-elles corrélées ? Vous pourrez utiliser la commande :

```
> abline(lm(Cal ~ Sul, data=eaux), col="red")
```

pour tracer la droite de régression linéaire. Est-ce cohérent avec la corrélation calculée ci-dessus ?

### 3.2 Entre une variable quantitative et une qualitative

- i. Représentez le boxplot de la variable `Nitrate` pour chaque modalité du facteur `Pays` grâce à la commande suivante :

```
> boxplot(Nitrate ~ Pays).
```

Interprétez les résultats.

- ii. Explorez les autres combinaisons de variables quantitatives avec les variables qualitatives afin de déterminer les variables fortement liées. Pensez à ne laisser qu'une synthèse de vos observations dans le rapport.

### 3.3 Entre deux variables qualitatives

- i. Analysez la table de contingence entre les deux variables qualitatives `Type` et `Nature` avec la commande `table()`.
- ii. Représentez-la graphiquement avec la fonction `mosaicplot`. Interprétez les résultats.

## 4 Préparation des données pour l'ACP

Lors du prochain TP, nous ferons l'analyse en composantes principales (ACP) des données `eaux` étudiées dans ce TP. Afin de préparer la prochaine séance, faites les modifications suivantes :

- i. Renommez les lignes du tableau `eaux` avec leurs acronymes (vous pourrez utiliser la fonction `row.names`).
- ii. Ne gardez dans `eaux` que les deux variables qualitatives `Type` et `Nature`, ainsi que toutes les variables quantitatives.
- iii. Enfin, enregistrez le jeu de données mis en forme dans le fichier `eauxModif.txt` :  

```
> write.table(eaux, file="eauxModif.txt",  
              row.names=TRUE, col.names=TRUE)
```
- iv. Enfin, vérifiez que le package `FactoMineR` est bien installé.