Université d'Orléans 2025–2026
UFR ST Miage M1

**Analyse de données**
**TP n°2** : Analyses de variables qualitatives et quantitatives

Ce TP utilise les données *StateFacts* du TP n°1 avec notamment le facteur qualitatif créé à partir de Diplome.

# 1 Variables qualitatives : tables de contingence

**(1)** Donner avec `table` les tables de fréquences des facteurs qualitatifs FD et Région. Idem pour la table des probabilités empiriques.

**(2)** Donner les représentations graphiques *barplots* associées : dans R il y a (au moins) 2 manières d'obtenir ce graphique ; donnez le barplot des effectifs $n_j$ et celui des probabilités empiriques $f_j = n_j/n$, $j = 1, \ldots, K$.

**(3)** Donner la table de contingence croisant facteurs Diplôme (`FD`) et Région. Interpréter

**(4)** Donner les tables des profils-ligne, puis colonne, de cette table de contingence. (voir `?prop.table`). Interpréter.

**(5)** Existe-il un effet Région sur le taux de diplôme donné par `FD` ? Donnez les représentations graphique associées à cette question (voir `?spineplot`). Que concluez-vous ?

# 2 Variables qualitatives : test d'indépendance du $\chi^2$

**(1)** Effectuez le test du $\chi^2$ pour $H_0$ : "`FD` et `Region` sont indépendantes" contre $H_1$ : "c'est faux". Concluez.

**(2)** Représentez la densité de la loi sous $H_0$, avec la valeur observée de la statistique de test et la région de rejet à 5%

**(3)** Le test ci-dessus a un problème de validité ; pourquoi ?
on peut tenter de le résoudre en recodant Diplôme en $k = 3$ classes égales en probabilité.

1. Effectuez ce recodage à l'aide de `cut`
2. appliquez à nouveau le test d'indépendance ; que concluez-vous ?
3. Si le problème de validité persiste, que proposez-vous ?

**(4)** Pour généraliser cette procédure, écrire une fonction (voir `?"function"`) qui :

1. prend comme arguments un vecteur de données $X$, le nombre de classes $k$, les noms des modalités
2. retourne un facteur qualitatif, recodage de $X$ en $k$ classes égales en probabilité, de modalités par défaut `{"1",...,"k"}`

# 3 Variables quantitatives

**Analyses préliminaires d'un jeu de données** C'est l'une des premières choses à faire lorsqu'on étudie un jeu de données ; il s'agit de :

1. détecter la présence de valeurs manquantes, aberrantes
2. étudier l'allure des distributions empiriques, à la recherche d'extrêmes, de formes gaussiennes ou au contraire non standards...

**(1)** Le jeu de données contient-t-il des valeurs manquantes ? aberrantes ? répondez en une seule commande (il ne s'agit pas de balayer le jeu de données visuellement !).

**(2)** Remplacez dans le `data.frame` l'observation de Population pour l'état de la ligne n°12 par manquant.

1. ré-exécutez les commandes de (1) pour visualiser l'existence d'un manquant
2. affichez l'observation (la ligne) pour laquelle la valeur de Population est manquante ; dans une vraie analyse c'est ainsi que l'on repère les observations à corriger, etc.
3. restorez le `data.frame` original pour la suite (attention au mécanisme `attach`)

**(3) Histogrammes :** c'est un moyen plus fin de détecter les valeurs aberrantes, et de visualiser les distributions empiriques. Affichez et interprétez les histogrammes :

1. par défaut pour les variables criminalité et analphabétisme
2. les mêmes normalisés pour la densité
3. en essayant d'autres choix automatiques de classes (voir `breaks =`)
4. pour toutes les variables quantitatives de cette table, en une seule commande en utilisant `apply`

**(4)** En utilisant une boucle `for (...) {...}`, affichez dans un unique graphique les mêmes histogrammes que ci-dessus (1.4), mais avec des titres utilisables. Que pouvez-vous conclure ?

**(5)** On peut remarquer qu'un état a une valeur de Revenu extrême par rapport à la distribution des revenu du reste de la population. Il est naturel de chercher à en savoir plus. En regardant les aides de `sort` et `order`, et en une seule commande, affichez l'observation (ligne) de cet état de plus fort revenu.

**(6)** Avec un code analogue à **(4)**, affichez les distributions boxplot des mêmes variables. Que remarquez-vous ?

# 4 Croisement d'un facteur avec une variable quantitative

**(1)** Le barplot d'un facteur avec une statistique numérique d'une variable est une des premières manières de croiser les deux types de variables. Affichez les barplots de Région avec moyennes puis écart-types de Diplôme, puis criminalité. A quel type de question ce résumé graphique répond-il ? Que concluez-vous ?

**(2)** Affichez les barplots de `FD` avec moyennes de Diplome. Ce résultat-ci est-il surprenant ?

**(3)** Mieux car plus informatif : la distributions boxplot d'une variable quantitative, ventilée par sous-populations associée à un facteur. Représentez les distributions boxplot de Criminalité par Région, voir `?boxplot`.

**(4)** En utilisant une boucle `for` comme précédemment, affichez dans un unique graphique les distributions boxplot des variables quantitatives de cette table par Région.

**(5)** Utilisez sur le même type de graphique les options `varwidth, notch` de boxplot.

## 5 Nuages de points – Scatterplots

**(1)** Visualisez l'effet des méthodes de `plot` pour les arguments suivants :

1. une variable quantitative
2. deux variables quantitatives
3. une variable quantitative et un facteur
4. un `data.frame`

**(2)** Quels nuages de points vous paraîssent les plus intéressant à représenter, au vu de l'étude de statistiques descriptives faite jusqu'ici ? représentez-les. Interpréter les graphiques obtenus.

**(3)** Scatterplots avec représentation d'un facteur

1. représentez le nuage de criminalité fonction d'analphabétisme avec labels individus
2. représentez le nuage de criminalité fonction d'analphabétisme avec labels des Régions (utilisez `RG` créé au TP1, pourquoi ?)
3. représentez le nuage de criminalité fonction d'analphabétisme avec labels individus et coloration par région.

Quel lien avec les études précédentes ?

**(4)** Chargez le package `ade4`, puis représentez quelques nuages pertinents avec labels individus et coloration et barycentres par Région (voir `?s.class`).

## 6 Analyse de $p$ variables quantitatives

**(1)** Les techniques présentées ci-dessus permettent d'analyser deux-à-deux des variables quantitatives. Une vision plus globale des liens linéaires entre $p$ variables quantitatives est obtenue par lma matrice des corrélations empiriques.

1. Donner la matrice de corrélation des variables qualitatives de cette table
2. R par défaut affiche 8 décimales. Affichez cette matrice de manière plus lisible
3. affichez uniquement les corrélations entre la variable à expliquer criminalité, et les autres variables

Même cette matrice devient difficile à utiiser lorsque $p$ devient grand. On passe alors aux méthodes de Data Mining comme l'ACP.