Année 2024-2025

Analyse de Données

Travaux Pratiques avec le logiciel R
TC2A
Sandie Ferrigno

TP 1

Initiation au logiciel R
Statistique descriptive et inférentielle

Initiation au logiciel R.

Exercice 1 : Pour commencer avec R.

1. Lancer le logiciel R. Le signe > qui apparaît dans la fenêtre R console signifie que R est prêt à travailler. A la suite de ce symbole, exécuter et commenter les commandes suivantes :

   $2 + 5$
   $sqrt(4)$
   $a <- 5$
   a
   $a = 3$
   a
   $120 : 155$

2. Si vous quittez R, une question vous est posée : sauver une image de la session ? Si vous répondez oui, le logiciel conserve tous les objets créés auparavant. Nous ne pourrons pas le vérifier dans le cadre de ce TP car nous n'avons pas les droits d'écriture nécessaires. Vous pourrez le vérifier si vous installez R sur votre ordinateur personnel.

3. Ouvrir un script dans R. Taper les commandes de la question 1. dans ce script et les exécuter. Enregister votre script.

Exercice 2 : Suites.

Les commandes suivantes permettent d'obtenir des suites de nombres qui peuvent être stockées dans des objets. Exécuter ces commandes, commenter les sorties et répondre aux questions posées.

1. suite$<-1 : 12$
   suite
2. suite1$<-seq(from = 1, to = 12, by = 1)$ (ou $seq(1, 12, 1)$)
   suite1
3. suite2$<-seq(1, 12, 2)$
   suite2
4. Construire une suite de nombres allant de 20 à 50 par pas de 5.

Exercice 3 : Vecteurs.

Les vecteurs sont des objets d'un même mode constitués d'éléments de type numérique, caractère, logique, vide. Exécuter les commandes suivantes, commenter les sorties et répondre aux questions posées.

1. Construire un vecteur que vous nommerez vecteur1 composé des éléments suivants : $5, 6, 1, 4, 5$. Afficher ce vecteur.

2. Construire un vecteur que vous nommerez vecteur2 composé des éléments suivants : bleu, vert, jaune. Afficher ce vecteur.

3. vecteur3$<-c(T, T, F, T, F, F)$
   vecteur3

4. mode(vecteur1)
   Quels sont les modes des objets vecteur2 et vecteur3 ?

5. length(vecteur1)
   Quelles sont les longueurs des objets vecteur2 et vecteur3 ?

6. is.vector(vecteur1)
   is.vector(suite2)

   Affichage d'un ou plusieurs éléments d'un vecteur. Répondre aux questions suivantes :

7. Afficher uniquement l'élément 2 de vecteur1.
8. Afficher les éléments 2 à 4 de vecteur1.

   Diverses manipulations sont possibles sur les vecteurs : concaténation, extraction, calculs, répétitions, légende et tri. Regarder ce que renvoient les commandes suivantes et répondre aux questions posées :

9. $x <-c(2, 3, 6, 10, 12)$
   $y <-c(1, 6, 7, 2, 1)$
   $z <-x + y$
   z
   $2 * x + 5$
   $(x + y)/2$

10. Extraire les éléments 2 et 4 du vecteur x.
    Supprimer les éléments 2 et 3 du vecteur x.
    Extraire du vecteur y les éléments strictement supérieurs à 4.

11. $x[3] <-35$
    x
    $y[y == 1] <-25$

y
Remplacer dans le vecteur x tous les éléments supérieurs ou égaux à 5 par 20.

12. $donnees <- c(1, 2, 3)$
$rep(x = donnees, times = 2)$ ou $rep(donnees, 2)$
Créer un vecteur contenant 50 fois le chiffre 1.
Créer un vecteur contenant 5 fois le mot MINES.

13. notes<−c(Anglais=12,Informatique=19,Maths=8)
notes
matieres<−c("Anglais","Informatique","Maths")
evaluation<−c(12,19,8)
evaluation
names(evaluation)<−matieres
evaluation

    Remarque : pour supprimer les noms, il suffit de taper la commande
    names(evaluation)<−NULL
    puis evaluation.

    Classer par ordre croissant (puis décroissant) les éléments de l'objet "evaluation".

Exercice 4 : Matrices.

Les matrices sont des objets mathématiques très utiles en Statistique, plus particulièrement en Analyse de données. Nous allons apprendre à les construire sous R et à les manipuler : affichage de certains éléments, calculs,....

1. matrice1<−matrix(1 :12,ncol=3)
matrice1
matrice2<−matrix(1 :12, ncol=3,byrow=TRUE)
matrice2
Quelle est la différence entre les objets matrice1 et matrice2 ?
Donner le nombre d'éléments de matrice1.
Donner les dimensions de matrice1.

2. Extraire l'élément de la deuxième ligne et la troisième colonne de matrice1.
Extraire la deuxième ligne de matrice1.
Extraire la troisième colonne de matrice1.
matrice1[,3,drop=F]

3. nrow(matrice1)
ncol(matrice1)
Ajouter la ligne composée des éléments 13, 14 et 15 à matrice1.
Ajouter la colonne composée des éléments 13, 14, 15 et 16 à matrice1.

4. matrix1<−matrix(1 :6,ncol=3)
   matrix1
   matrix2<−matrix(1 :12,ncol=4)
   matrix2
   Peut-on effectuer le produit matriciel de matrix1 par matrix2 ? Si oui, le calculer.
   Calculer la transposée de matrix1.
   matrix3<−matrix(1 :4,ncol=2)
   Calculer le déterminant de matrix3.
   La matrice matrix3 est-elle inversible ? Si oui, calculer son inverse.
   Diagonaliser matrix3. Vous donnerez les valeurs propres et vecteurs propres de cette matrice.

Exercice 5 : Les listes.

Une liste est une collection ordonnée d'objets, non nécessairement de même mode. Les éléments d'une liste peuvent donc être n'importe quel objet défini dans R. Exécuter les commandes suivantes, commenter les sorties et répondre aux questions posées.

1. $li <- list(num = 1:5, y = "couleur", a = T)$
   li
   $li\$num$
   $li\$a$
   Extraire le premier élément de la liste li.
   Extraire le troisième élément de la liste li.

2. $vec <- c(1, 2, 3)$
   $mat <- matrix(1, ncol = 2, nrow = 3)$
   $L <- list(vec, mat)$
   Extraire le premier élément de la liste L.
   Extraire le deuxième élément de la liste L.
   $L <- list(vecteur = vec, matrice = mat)$
   L
   $L[[2]]$ ou $> L\$matrice$

3. $m <- matrix(1:4, ncol = 2)$
   $dec <- eigen(m)$
   dec
   Extraire les valeurs propres de la liste dec.
   Extraire les vecteurs propres de la liste dec.

Exercice 6 : Les Data-frames 1.

Les data-frames (ou structures de données) sont des tableaux dont les colonnes peuvent être de différents types. Ce sont des listes particulières dont les composantes sont de même

longueur mais de mode différent. Ils sont les objets les plus importants de R, ils nous permettrons par la suite de faire des statistiques sous R. Exécuter, commenter les commandes suivantes et répondre aux questions posées.

$age <- c(17, 28, 64, 8, 25, 36)$
age
$sexe <- c("H", "F", "F", "H", "H", "F")$
$sexe$
$donnees <- data.frame(age, sexe)$
$donnees$
Extraire l'élément sur la troisième ligne et la première colonne du data-frame.
Extraire la quatrième ligne du data-frame.
Extraire la deuxième colonne du data-frame.
$donnees[[2]]$
$donnees\$sexe$
$names(donnees)$

Exercice 7 : Les Data-frames 2.

R est un ensemble de bibliothèques appelées packages. Chaque package contient des jeux de données spécifiques qui sont utilisés comme exemples. Le but de cet exercice est d'utiliser le jeu de données iris qui fait partie du package "base". Exécuter, commenter les commandes suivantes et répondre aux questions posées.

$iris$
Chercher des informations sur le fichier iris.
Extraire les individus 1 à 5 du jeu de données iris.
$names(iris)$
$length(iris)$
$dim(iris)$

Toujours à partir du jeu de données iris :

1. créer un nouveau jeu de données comportant uniquement les données de la modalité versicolor de la variable species et le nommer iris2 (on cherchera une autre solution que iris[51 :100,]).
2. trier par ordre décroissant les données de iris2 en fonction de la variable Sepal.Length (utiliser la fonction order).

Exercice 8 : Les Data-frames 3.

Quand les données sont plus volumineuses, nous utilisons un éditeur de texte pour saisir les données puis nous les transférons ensuite sous R.

Les données suivantes ont été saisies dans le fichier table1.txt :

53.5 160
74.4 172
52.6 151
88.6 163
49.2 169

1. Changer le répertoire de travail associé à R. Pour cela, exécuter la commande "getwd()". Cette commande permet de savoir quel est le répertoire que va utiliser le logiciel R pour faire appel à des fichiers externes. Pour modifier ce répertoire, taper la commande "setwd("U :/")" pour par exemple choisir "U :/" comme répertoire de référence. Vous pouvez alors exécuter à nouveau la commande "getwd()" pour vérifier que le changement de répertoire a bien été pris en compte.

   Remarque : Ce changement de répertoire est valable le temps de la session R. Cela signifie que vous devrez modifier le répertoire au début de chaque séance de TP.

2. Télécharger le fichier table1.txt et le placer dans le répertoire actuel de travail (celui de la question 1).

3. Créer le nouveau jeu de données *tab*1 dans R à partir du fichier table1.txt.

4. Lorsque le nom des variables est spécifié dans la première ligne du fichier texte de référence (voir table2.txt) il faut l'indiquer lors de l'importation du fichier sous R. Créer un nouveau jeu de données nommé tab2 à partir du fichier table2.txt.

5. Si dans le fichier texte les décimales sont notées , au lieu de . (comme c'est le cas dans table3.txt) il faut le spécifier lors de l'importation du fichier sous R. Créer un nouveau jeu de données nommé tab3 à partir du fichier table3.txt.

6. D'autres formats que le format texte sont utilisés pour stocker les données, notamment csv et excel. Créer un nouveau jeu de données nommé tab4 à partir du fichier table4.csv.

Exercice 9 : Les packages.

Il existe de nombreux packages R. Nous allons voir au travers de cet exercice un exemple d'utilisation. Le jeu de données Europe se trouve dans le package BioStatR.

1. Installer et charger le package BioStatR. Pour cela, aller dans Menu, Packages, Installer Packages. Puis, choisir un site mirroir du CRAN, par exemple France, Lyon 1. Choisir enfin le nom du package, ici BioStatR. Il suffit ensuite d'activer le package pour pouvoir l'utiliser le temps de votre session R. Pour cela, aller dans Menu, Packages, Charger le package et choisir BioStatR.

2. Afficher le jeu de données Europe.

3. Fermer puis réouvrir R. Taper la commande Europe. Que constatez-vous ?

Statistique descriptive et inférentielle.

Exercice 10 : Le jeu de données nommé chickwts fait partie du package base de R. Il fournit le poids de 71 poulets selon le type de nourriture intégrée. Nous allons donc travailler dans cet exercice avec 71 individus et 2 variables : une de type quantitatif et l'autre de type qualitatif.

1. On s'intéresse d'abord à la variable weight.
   — De quel type de variable s'agit-il ? Créer un vecteur contenant ses éléments.
   — Calculer la moyenne, la variance et l'écart-type de cette variable.
   — Construire l'histogramme de la distribution des poids. Interpréter ce graphique.
   — Construire un graphique de type Box-Plot de la distribution des poids. Interpréter ce graphique.
   — Calculer la médiane, le premier et le troisième quartile sur ces données. Interpréter ces résultats.
2. Regardons maintenant la variable feed.
   — De quel type de variable s'agit-il ? Créer un vecteur contenant ses éléments.
   — Etablir la liste des modalités de cette variable ainsi que les effectifs de chacune des modalités.
   — Dresser le tableau des fréquences. Interpréter ces résultats.
   — Représenter cette variable à l'aide d'un graphique en secteurs ou camembert. Interpréter ce graphique.
   — Représenter cette variable à l'aide d'un graphique en bâtons. Interpréter ce graphique.
3. On souhaite savoir si le type de nourriture ingérée influe sur le poids du poulet. Représenter cette information sur un graphique. Lequel vous paraît le plus adapté ? Interpréter ce graphique.

Exercice 11 : Indice de Quételet.

Nous avons recueilli un certain nombre d'informations sur le sexe, le poids (en kg) et la taille (en cm) d'un échantillon d'hommes et de femmes. Elles sont stockées dans la table quetelet.txt.

1. Importer le fichier de données sous R.
2. Créer trois vecteurs taille, poids et sexe correspondants aux données de la table.
3. Quel est le nombre d'individus dans l'échantillon.
4. Pour les deux variables quantitatives calculer quelques indicateurs de tendance centrale.

5. Nous souhaitons calculer et étudier l'indice de Quételet à partir des données précédentes. La formule de cet indice est la suivante :

$$\text{indice} = \frac{\text{Poids (en kg)}}{(\text{Taille (en m)})^2}.$$

Cet indice permet de mesurer la corpulence de l'homme adulte. Davenport a établi la classification suivante : très maigre (moins de 18,1), maigre (de 18,1 à 21,4), moyen (de 21,5 à 25,6), corpulent (de 25,7 à 30,4) et obèse (30,5 et plus).

— Calculer les paramètres statistiques élémentaires de cette nouvelle variable sur l'ensemble des individus puis en fonction du sexe.
— Construire l'histogramme de cette nouvelle variable sur l'ensemble des individus. Vous préciserez les classes au logiciel.
— Construire les deux histogrammes des hommes et des femmes.

Exercice 12 : Les poids à la naissance (en kg) de deux échantillons de veaux de deux races différentes sont les suivants :

Race Parthenaise	53	49	40	48	43	42	43	46	42	43	38	40	50	44
Race Charolaise	46	46	48	38	42	42	40	53	55	41	47	30		

1. Calculer la moyenne et l'écart type de chacun de ces échantillons.
2. Au risque a de 5%, peut-on considérer que les variances des deux populations sont égales ?
3. Peut-on conclure, au vu de ces échantillons et au risque de 5%, que les poids moyens à la naissance des deux races sont différents ?

Exercice 13 : On donne ci-après les pourcentages de matière grasse dans un aliment, déterminés sur 10 échantillons par deux méthodes d'analyse différentes A et B.

Echantillon	1	2	3	4	5	6	7	8	9	10
Méthode A	24.6	25.3	25.3	25.6	25.6	25.9	26	27	27.3	27.7
Méthode B	24.9	25.6	25.8	26.2	26.1	26.7	26.3	26.9	28.4	27.1

Comparer ces deux méthodes.

Exercice 14 : Le tableau suivant donne la répartition de 10000 personnes en fonction de leur groupe sanguin et de leur facteur Rhésus.

	O	A	B	AB
Rh_+	3535	3870	1000	158
Rh_-	665	630	100	42

Les facteurs groupe sanguin et rhésus sont-ils indépendants ?

TP 2

Régression Linéaire

Exercice 1 : Le fichier **freinage.txt** contient les variables vitesse et distance. Elles concernent les données relatives aux distances de freinage jusqu'à l'arrêt complet d'une voiture lancée à différentes vitesses.

1. Créer la table de données freinage à partir du fichier texte freinage.txt.
2. Tracer le nuage de points de "distance" en fonction de "vitesse".
3. Calculer le coefficient de corrélation entre ces deux variables. Est-il significatif au risque de 5% ?
4. Ajuster un modèle de régression linéaire pour ces distances en fonction des vitesses.
5. Donner l'équation de la droite de régression estimée et tracer cette droite.
6. Donner une estimation de l'écart-type résiduel.
7. Combien vaut R^2 ? Interpréter cette valeur.
8. Tester la significativité des paramètres de la régression.
9. Donner les IC des paramètres de la régression.
10. Vérifier les hypothèses de cette modélisation sur le graphe des résidus. Que pouvez-vous en conclure ?
11. Trouver une transformation de la variable explicative qui permettrait de contourner le problème concernant les résidus.
12. Reprendre les points précédents en ajustant ce nouveau modèle. Conclusion ?
13. Donner une estimation de la distance d'arrêt pour une vitesse de 60km/h.
14. Donner les IC à 95% de la moyenne d'une observation et d'une observation pour une vitesse fixée à 60km/h.

Exercice 2 : Le fichier **bacterie.txt** présente la décroissance d'une population de bactéries dans un milieu soumis à une exposition aux rayons X pendant 15 intervalles de temps de 6 minutes.

1. On veut analyser l'évolution de la variable nombre (nombre de centaines de bactéries survivantes) en fonction de la variable temps. Tester le modèle linéaire nombre/temps et examiner le graphe des résidus. Vous pourrez pour cela vous aider des questions 1 à 10 de l'exercice 1.

2. Le modèle précédent est manifestement inadéquat. Si l'on en croit les biologistes, la décroissance de la population peut s'écrire : $n_t = n_0 e^{\beta t}$, $t \geq 0$. Le paramètre n_0 est le nombre (centaines) de bactéries au départ de l'expérience et β est le taux de "destruction". Comment modifier le modèle pour tenir compte de cette information et de ce que vous avez constaté sur le graphe des résidus précédent ?

Exercice 3 : En physiologie, un moyen pour étudier la forme physique est de savoir à quelle vitesse le corps peut absorber et utiliser l'oxygène. Des sujets ont participé à un exercice prédéterminé qui consiste en une course à pied de 2km400. On a alors enregistré des mesures de leur consommation d'oxygène ainsi que d'autres variables continues telles que l'âge, le pouls et le poids. Des chercheurs se sont intéressés à comment pourrait on prédire la consommation d'oxygène à partir de ces variables. Le jeu de données **b-fitness** est issu de Rawlings (1998) et contient les variables suivantes :

Name : Nom du participant.
Gender : Sexe du participant.
Runtime : Temps en minutes pour parcourir 2km400.
Age : Age (en années) du participant.
Weight : Poids (en kg) du participant.
Oxygen-Consumption : Mesure de la capacité à utiliser l'oxygène qui se trouve dans le sang.
Run-Pulse : Pouls à la fin de la course.
Rest-Pulse : Pouls au repos.
Maximum-Pulse : Pouls maximum durant la course.
Performance : Note globale de la forme physique.

1. Tracer les nuages de points de la variable "Oxygen-Consumption" via chacune des variables quantitatives données dans l'énoncé. Interprétation ?
2. Calculer les coefficients de corrélation entre la variable "Oxygen-Consumption" et chacune des autres variables quantitatives données dans l'énoncé. Interprétation ?
3. Calculer les coefficients de corrélation entre les variables quantitatives autres que "Oxygen-Consumption" utilisées dans la question 2. Interprétation ?
4. Effectuer une régression multiple de "Oxygen-Consumption" en prenant toutes les autres variables quantitatives comme régresseurs. Interpréter les divers résultats.
5. Effectuer une sélection de modèle dans le cadre de la régression multiple de "Oxygen-Consumption" sur les autres variables quantitatives données dans l'énoncé. Commenter.

6. Vérifier les hypothèses relatives au modèle retenu.

Exercice 4 : L'enquête "**Prix et Salaires**", que publie UBS (Union des Banques Suisses) tous les trois ans, brosse un tableau mondial des prix des produits et services, des salaires, des retenues à la source et des heures de travail, et du pouvoir d'achat qui en découle, dans 71 villes sur tous les continents. Nous avons sélectionné pour notre étude de cas, 12 variables de "prix", une variable de "salaire" et la variable "villes".

Le problème est d'étudier le salaire horaire net moyen en euros (variable S_H_NET) en fonction des 12 variables suivantes de prix que nous avons retenues, renseignées pour 69 villes de l'enquête (tous les prix sont en euros).

- **ALIM** : Prix d'un panier pondéré avec 39 denrées alimentaires. Dépense mensuelle de famille occidentale moyenne.
- **VET_F** : Garde-robe complète pour dame composée d'un tailleur (deux pièces), d'une veste, d'une jupe, de collants et d'une paire de chaussures de ville à la mode.
- **VET_H** : Garde-robe complète pour homme composée d'un costume, d'un blazer/d'une veste, d'une chemise, d'un jean, de chaussettes et d'une paire de chaussures de ville.
- **EQUIP** : Coût d'un panier avec : un réfrigérateur, un téléviseur couleur, un appareil photo, un fer à repasser à vapeur, un aspirateur, une poêle, un sèche-cheveux et un PC.
- **MEUBL** : Loyers d'appartements construits après 1980 (4 pièces, cuisine, bain, garage, y compris les charges), correspondant au "standing" d'un cadre moyen et situé dans un secteur privilégié par un tel cadre.
- **APPART** : Loyers d'appartements construits après 1980 (3 pièces, cuisine, bain, sans garage, y compris les charges) répondant dans l'ensemble aux exigences de confort locales, situés à proximité du centre ville.
- **TRANSP** : Prix pour un trajet de 10km ou de 6 miles environ ou de 10 stations au minimum.
- **TAXI** : Prix pour une course de 5km ou de 3 miles effectuée de jour dans le périmètre urbain, service compris.
- **AUTO** : Prix d'une voiture de classe moyenne la plus vendue (toutes taxes comprises), 5 portes, équipement de série.
- **RESTO** : Prix d'un dîner (trois plats avec entrée, plat principal, dessert, sans boisson), service compris, dans un bon restaurant.
- **HOTEL** : Prix d'une chambre à deux lits, avec bain et WC, y compris le petit

déjeuner pour deux personnes dans un hôtel *** de première classe (catégorie internationale) ou de standard moyen.

— **SERVICES** : Panier de 27 biens et services non transférables : coupe de cheveux, pressing, facture de téléphone, billet de cinéma, connection internet DSL, frais d'inscription pour différents cours et billets pour des activités de loisirs.

— **S_H_NET** : Salaire horaire net en euros (rémunération effective recensée dans 14 professions, compte tenu du temps de travail, des jours fériés et des vacances. Pondération selon la représentativité des professions.

— **VILLES** : villes des divers continents participantes à l'étude.

TP 3

Analyse en composantes principales

Exercice 1 : ACP sur les données **Jus d'orange** (Cornillon et al., Statistiques avec R, 2012).

Les données ont été recueillies dans le cadre de travaux d'étudiants d'Agrocampus à Rennes. Six jus d'orange ont été évalués par un jury d'étudiants selon sept variables sensorielles : intensité de l'odeur, typicité de l'odeur, caractère pulpeux, intensité du goût, caractère acide, caractère amer et caractère sucré. Ce sont les moyennes des évaluations du jury qui se trouvent dans le tableau de données. En plus de ces descripteurs sensoriels, on dispose de variables physico-chimiques telles que le glucose, le fructose, le saccharose, le pouvoir sucrant, le PH, l'acide citrique et la vitamine C. Le but de l'exercice est de décrire les jus d'orange à partir de leur seul profil sensoriel. Cette problématique pourra être enrichie en reliant les dimensions sensorielles aux variables physico-chimiques. Dans cet exemple, on a également introduit la variable conditionnement qui prend les modalités "ambiant" ou "frais" ainsi que la variable origine des jus de fruit qui prend les modalités "Floride" ou "Autre".

1. Importer le jeu de données sous R.
2. Donner quelques statistiques descriptives sur les variables d'intérêt.
3. Donner la matrice des corrélations entre les diverses variables de l'étude. Interpréter.

   Nous allons à présent réaliser une ACP sur les jus d'orange et les variables sensorielles.
4. Donner les valeurs propres de la matrice des corrélations. Interpréter les résultats et notamment déterminer le nombre d'axes intéressants pour l'étude.
5. Donner les coordonnées, les corrélations, les cosinus carrés et les contributions des variables dans un premier plan factoriel qui vous semble être celui contenant le plus d'information.
6. Donner les coordonnées, les cosinus carrés et les contributions des individus dans un premier plan factoriel qui vous semble être celui contenant le plus d'information.
7. Représenter graphiquement les variables et les individus dans ce premier plan factoriel. Interpréter vos résultats.

8. Reprendre les questions 5, 6 et 7 si d'autres études selon d'autres axes vous paraissent pertinentes.

Exercice 2 : ACP sur les données **Décathlon** (Cornillon et al., Statistiques avec R, 2012).

Le jeu de données décathlon concerne les résultats aux épreuves du décathlon lors de deux compétitions d'athlétisme qui ont eu lieu à un mois d'intervalle : les jeux olympiques d'Athènes (23 et 24 août 2004) et le Décastar (25 et 26 septembre 2004). Le premier jour, les athlètes participent à 5 épreuves (100m, longueur, poids, hauteur, 400m) et le deuxième jour, aux 5 épreuves restantes (110m haies, disque, perche, javelot, 1500m). Dans le tableau de données, on trouvera pour chaque athlète, outre ses performances à chacune des 10 épreuves, son classement final, son nombre de points final et la compétition à laquelle il a participé.

L'objectif de cet exercice est de déterminer des profils de performances similaires : y-a-t-il des athlètes plus endurants, plus explosifs,.... ? Certaines épreuves se ressemblent-elles ? Si un athlète est performant pour une épreuve, est-il plutôt performant pour une autre ? Pour répondre à ces questions, réaliser une ACP sur ce jeu de données. On prendra soin de n'utiliser que les variables qui vous sembleront pertinentes pour l'étude. Vous pourrez vous aider des questions de l'exercice 1 pour mener cette étude.

Exercice 3 : ACP sur les données **Meteo35villes** (Husson et al., Analyse de Données avec R, 2016).

On s'intéresse ici au climat des différents pays d'Europe. Dans le jeu de données "meteo35villes", on a recueilli les 12 températures mensuelles moyennes (en degrès Celsius) pour 35 grandes villes Européennes. En plus des températures mensuelles, on donne, pour chaque ville, la température annuelle moyenne, l'amplitude thermique (différence entre la moyenne mensuelle maximum et la moyenne mensuelle minimum d'une ville). On donne également deux variables quantitatives de positionnement (la longitude et la latitude) ainsi qu'une variable qualitative, region (appartenance à une région de l'Europe, variable à 4 modalités : Europe du Nord, du sud, de l'Est, de l'Ouest). On souhaite appréhender la variabilité des températures mensuelles d'un pays à l'autre de façon multidimensionnelle. Réaliser une ACP sur ce jeu de données. On prendra soin de n'utiliser que les variables qui vous sembleront pertinentes pour l'étude. Vous pourrez vous aider des questions de l'exercice 1 pour mener cette étude.

TP 4

Analyse Factorielle des Correspondances Simples et Multiples

Exercice 1 : AFC sur les données **Universités** (Cornillon et al., Statistiques avec R, 2012).

Le jeu de données « Universités » représente le nombre d'étudiants des universités françaises par discipline et par cursus selon le sexe lors de l'année 2007-2008. Il s'agit d'un tableau qui croise deux variables qualitatives « Discipline » et « Niveau-sexe ». Nous disposons de plus, par discipline, du nombre total d'étudiants par niveau, par sexe et du total global. Le but de cette étude est d'avoir une image de l'université. Quelles sont les disciplines pour lesquelles le profil des étudiants est le même ? Quelles sont les disciplines privilégiées par les femmes ? (respectivement par les hommes ?) Quelles sont les disciplines pour lesquelles les études sont plus longues ?

Exercice 2 : Enquête sur les **séjours-vacances** des français (Saporta, 2006).

Le jeu de données « Vacances » provient de l'enquête sur les vacances des français en 1999 publiée par l'INSEE en mai 2002. Elle décrit la répartition des séjours selon la catégorie socio-professionnelle du chef de famille (CSP) et le mode d'hébergement. Les deux variables étudiées sont les suivantes :

CSP : agriculteurs, artisans, cadres et profession intellectuelles supérieures, professions intermédiaires, employés, ouvriers, retraités, autres inactifs.

Hébergement : hôtel, location, résidence secondaire, résidence principale parents amis, résidence secondaire parents amis, tente, caravane, auberge de jeunesse, village vacance.

La taille de l'échantillon est de 18532.
Existe-t-il un lien entre la catégorie socio-professionnelle (à 8 modalités) et le mode d'hébergement (à 9 modalités) des français ?

Exercice 3 : AFCM sur les données **Crédit** (Cornillon et al., Statistiques avec R, 2012).

Le jeu de données contient 66 clients ayant souscrit un crédit à la consommation dans un organisme de crédit. Les 11 variables qualitatives et les modalités associées à cet exemple sont les suivantes :

Marché : rénovation d'un bien, voiture, scooter, moto, mobilier-ameublement. Cette variable indique le bien pour lequel les clients ont réalisé un emprunt.

Apport : oui, non. Cette variable indique si les clients possèdent un apport personnel avant de réaliser l'emprunt. Un apport personnel représente une garantie pour l'organisme de crédit.

Impayé : 0, 1 ou 2, 3 et plus. Cette variable indique le nombre d'échéances impayées par le client.

Taux d'endettement : 1 (faible), 2, 3, 4 (fort). Cette variable indique le niveau d'endettement du client. Le taux d'endettement est calculé comme le rapport entre les charges (ensemble des dépenses) et le revenu. Ce taux a été discrétisé en 4 classes.

Assurance : sans assurance, AID (assurance invalidité et décès), AID + Chômage, Senior (pour les plus de 60 ans). Cette variable indique le type d'assurance à laquelle le client a souscrit.

Famille : union libre (concubinage), marié, veuf, célibataire, divorcé.

Enfants à charge : 0, 1, 2, 3, 4 et plus.

Logement : propriétaire, accédant à la propriété, locataire, logé par la famille, logé par l'employeur.

Profession : ouvrier non qualifié, ouvrier qualifié, retraité, cadre moyen, cadre supérieur.

Intitulé : M, Mme, Melle.

Age : 20 (18 à 29 ans), 30 (30 à 39 ans), 40 (40 à 49 ans), 50 (50 à 59 ans), 60 ans et plus.

Le but de cette étude est de caractériser la clientèle de l'organisme de crédit. Nous voulons mettre d'abord en évidence différents profils de comportements bancaires. Nous voulons ensuite étudier la liaison entre le signalétique (CSP, âge, etc.) et les principaux facteurs de variabilité des profils de comportements bancaires (c'est à dire caractériser les clients aux comportements particuliers).

TP 5

Classification hiérarchique

Exercice 1 : Classification sur les données **Fromages** (Chavent, page internet de cours).

Le jeu de données fromage contient des informations concernant 29 fromages. Ceux-ci sont décrits par 9 variables continues : calories, sodium, calcium, lipides, retinol, folates, proteines, cholesterol et magnesium.

Réaliser une classification hiérarchique ascendante de cet ensemble de fromages décrits par leurs propriétés nutritives. L'objectif est d'identifier des groupes de fromages homogènes, partageant des caractéristiques similaires.

Exercice 2 : Le jeu de données **autos2005** présente un échantillon de 40 voitures décrites par les variables suivantes : puissance, cylindrée, vitesse (maximale), longueur, largeur, hauteur, poids, coffre, réservoir, consommation, CO2, prix.

Le problème initial est de modéliser le prix des voitures en fonction de ces variables. Commencer par réaliser une ACP sur l'ensemble des variables de ce jeu de données (sauf le prix), ce qui permettra d'identifier des groupes indépendants de variables. Cela sera utile pour réaliser une régression du prix en fonction des variables explicatives tout en s'affranchissant des problèmes éventuels de colinéarité entre les variables explicatives.

Compléter cette ACP par une classification, toujours en excluant la variable prix des données. Comparer les différentes moyennes des prix de chacun des groupes obtenus. Terminer l'analyse par la projection des classes obtenues dans le plan factoriel 1-2 de l'ACP.

Exercice 3 : Le jeu de données **météo35villes** (déjà étudié lors du TP sur l'ACP) présente les 12 températures mensuelles moyennes pour 35 grandes villes Européennes. On y trouve également les variables moyenne, amplitude, latitude, longitude et région. Réaliser une classification sur ce jeu de données. Vous terminerez l'analyse par la projection des classes obtenues sur le plan factoriel 1-2 de l'ACP.

TP 6
Analyse discriminante

Exercice 1 : Analyse discriminante sur les données **Poissons** (Chavent, page internet de cours).

On dispose ici de 23 données concernant des poissons qui sont répartis en trois groupes selon leur site de pêche (site 1, site 2, site 3). La variable qualitative à expliquer Y est la variable site qui possède 3 modalités (site 1, site 2, site 3). Sur ces 23 poissons, on a mesuré les $p = 14$ variables quantitatives suivantes : YEU (radioactivité des yeux), BR (radioactivité des branchies), OP (radioactivité des opercules), NAG (radioactivité des nageoires), FOI (radioactivité du foie), TUB (radioactivité du tube digestif), EC (radioactivité des écailles), MUS (radioactivité des muscles), POI (poids), LON (longueur), LART (largeur de la tête), LAR (largeur), LARM (largeur du museau), DYEU (diamètre des yeux).

Le but de cet exercice est de pouvoir décrire Y ou encore expliquer l'appartenance à un site de pêche en fonction de ces 14 variables explicatives.

Exercice 2 : Analyse discriminante sur les données **Ronfle** (Cornillon et al., Statistiques avec R, 2012)

Dans le cadre d'une étude de la population angevine, le CHU d'Angers s'est intéressé à la propension à ronfler d'hommes et de femmes. Le fichier de données contient un échantillon de 100 patients. Les variables considérées sont :

- age (en années)
- poids (en kg)
- taille (en cm)
- alcool : nombre de verres bus par jour (en équivalent verre de vin rouge)
- sexe : sexe de la personne (F=femme, H=homme)
- ronfle : diagnostic de ronflement (O=ronfle, N=ne ronfle pas)
- taba : comportement au niveau du tabac (O=fumeur, N=non fumeur)

Le but de cette étude est d'essayer d'expliquer le ronflement (variable ronfle) par les six autres variables présentées ci-dessus. Pour cela, après avoir importé les données, vous construirez le modèle, estimerez le taux de mauvais classement et ferez de la prévision.

Exercice 3 : Nous souhaitons prédire le type (kirsch, mirabelle ou poire) d'un liquide alcoolisé contenu dans un verre selon sa composition. Nous disposons de 6 variables concernant sa composition (butanol, méthanol,.....). Nous disposons également de 77 observations

qui sont contenues dans le fichier alcool.csv. Réaliser une analyse discriminante à partir de ces informations.