LE LOGICIEL R COMME OUTIL D'INITIATION À LA STATISTIQUE DESCRIPTIVE : ENQUÊTE SUR LES DÉPENSES DES MÉNAGES

Hubert RAYMONDAUD[1]

TITLE

Introduction to descriptive statistics with the Software R: Household expenditure statistics

RÉSUMÉ

Le logiciel **R**, libre et gratuit, est un outil privilégié pour l'apprentissage de l'analyse exploratoire des données. En plus de sa gamme étendue de méthodes et de ses graphiques de qualité, il comporte un langage de commandes qui permet à l'utilisateur de se familiariser avec les techniques de statistique descriptive et de les approfondir ; grâce à lui, on peut facilement analyser de grands tableaux de données en tout ou en partie. **R** est présenté ici dans le cadre d'un travail pratiqué avec des étudiants du brevet de technicien supérieur agricole.

Mots-clés : *logiciel R, statistique descriptive, traitement des données.*

ABSTRACT

The free, open-source statistical software **R** is a primary tool for education in exploratory data analysis. In addition to a broad spectrum of methods and quality graphics, it features a command language that allows the user to gain familiarity with, and a thorough understanding of, descriptive statistical techniques; analyses can easily be performed with it, both on large data sets and subsets thereof. **R** is described here in the context of an experiment carried out with students from the higher agricultural technician certificate program.

Keywords: *software R, descriptive statistics, data analysis.*

1 Introduction

Cet enseignement avec **R** a été développé depuis trois ans avec des classes de BTSA[2], formation en deux ans après le baccalauréat pour le traitement et l'analyse des données issues de stages, collectifs ou individuels, réalisés dans le cadre du diagnostic de territoire et de la conception de projets de services en espace rural (BTSA SER).

Il est la transformation d'un enseignement fait depuis plus de trente ans, dans un cadre similaire, avec les divers outils logiciels dont je pouvais disposer dans les centres de formation et lycées où j'enseignais et dont l'apprentissage était compatible avec le niveau des étudiants et le temps dont ils disposaient pour ce travail.

Le diagnostic de territoire nécessite de traiter et d'analyser d'une part des données issues des bases de l'INSEE (portraits de zones[3]) et d'autre part les données issues d'enquêtes en relation avec les projets à développer.

[1] LEGTA Louis Giraud 84200 Carpentras-Serres, hubert.raymondaud@educagri.fr
[2] Brevet de Technicien Supérieur Agricole
[3] http://www.insee.fr/fr/bases-de-donnees/default.asp?page=statistiques-locales.htm

Quelle que soit l'origine des étudiants de BTSA de cette filière, j'ai toujours constaté qu'ils sont complètement démunis quand il s'agit de réinvestir les méthodes simples de la statistique descriptive, pourtant au programme des classes du collège et du lycée. À cette difficulté s'ajoute celle de la maîtrise de l'outil logiciel, tant pour la gestion des fichiers de données que pour la mise en œuvre des méthodes statistiques.

La séquence ici présentée, positionnée en début de formation, est organisée en trois parties associant des objectifs de réinvestissement de méthodes de statistique descriptive, des objectifs d'apprentissage des procédures **R** correspondantes et des objectifs d'interprétation des résultats obtenus et de rédaction d'un rapport de synthèse.

Les méthodes de statistique descriptive mises en œuvre sont celles d'une première exploration graphique et numérique permettant de caractériser l'échantillon. Cette première étape est très importante car elle permet de bien comprendre le profil des participants à l'enquête mais aussi les associations entre variables. Il est donc crucial de savoir bien utiliser les différents outils graphiques et numériques de l'analyse descriptive.

Les croisements de deux variables qualitatives (tableaux d'effectifs), d'une variable qualitative avec une variable quantitative (comparaisons de groupes), de deux variables quantitatives (nuages de points), et avec une variable qualitative (groupes de nuages), ne sont abordés que très partiellement au lycée, alors même qu'ils fournissent la plus grande partie des résultats de l'exploration des données nécessaires à une première analyse. C'est aussi à cette occasion que **R** révèle sa spécificité et toute son efficacité par rapport à l'utilisation d'un tableur, peu adapté à ce type de traitement, avec ce type de public.

En effet, bien que l'on puisse faire des nuages de points et des tableaux croisés d'effectifs avec les tableurs actuels, cela nécessite, pour chaque nouveau graphique ou tableau, de réorganiser les données, de sélectionner les lignes et colonnes correspondantes, et ensuite de juxtaposer manuellement les graphiques, ce qui rend les explorations longues et fastidieuses, difficultés difficilement surmontables dans le temps réduit dont on dispose avec des élèves peu habitués à gérer des données même de petite taille.

On ne rencontre pas ces difficultés avec **R** qui possède un langage permettant de travailler directement avec des groupes d'individus ou variables. Les graphiques peuvent être juxtaposés ou superposés au moment de leur création, dans des fenêtres particulières.

2 Contexte de l'exploration

2.1 Présentation du fichier de données

Le fichier **"HabitConso.csv"** contient un extrait de 205 individus statistiques (ménages) et 6 variables.

Les données sont extraites et adaptées à partir d'une enquête faite en 2000 dans le cadre d'un diagnostic de territoire de la communauté de communes de la COVE (Carpentras), pour un projet d'ouverture d'une maison de pays permettant la valorisation du patrimoine historique et naturel et la commercialisation de produits du terroir. Les étudiants de BTSA ont interrogé les personnes sortant d'une vingtaine de magasins alimentaires, en juillet, sur tout le territoire de la COVE.

Le fichier au format ".csv" (texte) permet l'importation directe dans **R**, quel que soit le format d'origine du fichier.

Les variables retenues sont :

RegionDomicile (qualitative)	Région du domicile des personnes interrogées
CSP (qualitative)	Catégorie socio-professionnelle
BudgetMensuelAlimentaire (quantitative)	Dépense mensuelle d'alimentation
BudgetMensuelFruits (quantitative)	Dépense mensuelle destinée aux fruits
RevenuMensuel (quantitative)	Revenu mensuel du ménage
BudgetAnnuelLoisir (quantitative)	Dépense annuelle destinée aux vacances

2.2 Pourquoi choisir l'utilisation du logiciel R ?

Mon choix du logiciel **R** relève de plusieurs considérations parmi lesquelles : la gratuité ; l'utilisation de plus en plus répandue dans l'enseignement supérieur et les organismes de recherche (INRA, INSERM, CNRS...) ; un langage de programmation interactif facile à apprendre, permettant aussi bien la mise en œuvre des méthodes de la description statistique et de l'inférence, même les plus récentes, que la programmation de simulations probabilistes simples ou complexes ; une bibliothèque de fonctions très fournie, rassemblées dans des "packages" et proposées sur internet par une communauté de développeurs, spécialistes des méthodes qu'ils proposent ; la possibilité de construire ses propres fonctions ; des graphiques d'une grande qualité et d'une grande variété ; des outils mathématiques comme le calcul matriciel, l'intégration numérique, l'optimisation...

J'utilise **R** avec les élèves en leur proposant un document comprenant les commandes à saisir pour réaliser les analyses demandées. Cela rend son utilisation rapide et facile, d'autant plus que l'on peut disposer d'un éditeur à coloration syntaxique (**Tinn-R**) qui facilite la lecture et l'écriture des lignes de commande. Une séance de deux heures est ainsi suffisante pour réaliser les premières analyses exploratoires. Au bout de deux séances, certains élèves écrivent déjà des procédures simples. Un traitement de cette enquête, au niveau d'exigence pour cette filière, se fait en 4 séances de deux heures.

La littérature française traitant de **R** est abondante. Les principales références que j'ai utilisées et que je conseille sont présentées dans la bibliographie en fin d'article.

3 Étapes de l'exploration

3.1 Importation et contrôle des données

La première étape consiste, comme toujours, à importer des données dans **R**, sous forme d'un objet **R** appelé **"data.frame"** qui est en fait un tableau de données individus × variables. On pourra ensuite commencer les descriptions statistiques. Les principaux objets de **R** sont les vecteurs (suites numériques ou alphabétiques indicées), les matrices (tableaux de nombres), les listes (collections d'objets de différents types), les "data.frame" qui sont les classiques tableaux de données individus × variables.

L'importation des données doit toujours être suivie du contrôle des données importées, à l'aide de quelques outils de **R** :

- la commande **`setwd()`** permet de fixer le chemin d'accès au dossier contenant le fichier à importer ;
- la commande **`read.table()`** importe le fichier ".csv" sous la forme d'un tableau de données ("data.frame") que l'on nomme dans cet exemple "habit" ;
- **`attach()`** permet l'utilisation de "habit" par défaut, **`names()`** liste le nom de toutes les variables du tableau importé ;
- **`summary()`** fait un résumé de toutes les variables du tableau ;
- **`habit[1:3,]`** extrait les données de toutes les variables des 3 premières lignes du tableau et les affiche.

Ci-dessous, dans les encadrés de présentation des lignes de commandes **R** et des résultats obtenus, les fonctions et le nom des paramètres sont de couleur rouge sombre, leurs arguments sont en orange, les résultats en vert.

```
setwd("chemin d'accès au dossier")
habit <- read.table("HabitConso.csv", sep = ";", header = T, dec = ",")
attach(habit)
names(habit)
  [1] "RegionDomicile"           "CSP"
  [3] "BudgetMensuelAlimentaire" "BudgetMensuelFruits"
  [5] "RevenuMensuel"            "BudgetAnnuelLoisir"
summary(habit)
   RegionDomicile         CSP     BudgetMensuelAlimentaire BudgetMensuelFruits RevenuMensuel  BudgetAnnuelLoisir
   CENTRE:56      COMMERCANT :42   Min.   : 257.0           Min.   :  6.0       Min.   : 520   Min.   :  85.0
   EST   :30      ENSEIGNANT :55   1st Qu.: 874.0           1st Qu.: 26.0       1st Qu.:1759   1st Qu.: 407.0
   NORD  :58      FONCTERRITA:47   Median : 973.0           Median : 43.0       Median :2732   Median : 735.0
   OUEST :26      LIBERAL    :28   Mean   : 929.5           Mean   : 47.6       Mean   :2698   Mean   : 890.5
   SUD   :35      OUVRIER    :33   3rd Qu.:1057.0           3rd Qu.: 59.0       3rd Qu.:3515   3rd Qu.:1303.0
                                   Max.   :1313.0           Max.   :138.0       Max.   :4925   Max.   :2005.0
habit[1:3,]
    RegionDomicile         CSP BudgetMensuelAlimentaire BudgetMensuelFruits RevenuMensuel BudgetAnnuelLoisir
  1            SUD FONCTERRITA                     1039                  24         3548               1664
  2          NORD     OUVRIER                      761                  86         1587                311
  3            SUD     LIBERAL                      973                 138         4773                735
  ...
```

3.2 Principaux traitements graphiques et résumés numériques

Les vérifications étant faites, on peut commencer l'analyse descriptive selon les phases suivantes :

1. la description de **chacune des variables qualitatives**, en utilisant les tableaux d'effectifs des tris à plat et les graphiques adéquats ;
2. la description des **relations entre les variables qualitatives deux à deux**, par les tableaux d'effectifs des tris croisés et les graphiques correspondants ;
3. la description des **associations entre des variables qualitatives et les variables quantitatives**, en réalisant des séries de graphiques juxtaposés permettant de repérer des groupes particuliers ;
4. la description **des relations entre les variables quantitatives deux à deux**, en représentant les nuages de points et les résumés numériques correspondant aux modèles d'ajustement choisis. On pourra compléter cela en réalisant le croisement de deux variables quantitatives par une variable qualitative, par exemple en utilisant des couleurs identifiant, sur les points du nuage, les modalités de la variable qualitative.

Cela permet de mettre en évidence d'éventuelles structures, découlant de l'effet des variables qualitatives.

3.2.1 Tris à plat de variables qualitatives : exemple de *RegionDomicile* et de *CSP*

Il s'agit donc simplement de dénombrer les effectifs de chaque modalité des variables pour construire les tableaux des effectifs et les illustrer.

Le logiciel permet, avec la fonction `plot(NomDeVariable,...)`, d'obtenir directement des diagrammes en barres sans être obligé de passer par la production du tableau des effectifs comme c'est le cas dans un tableur.

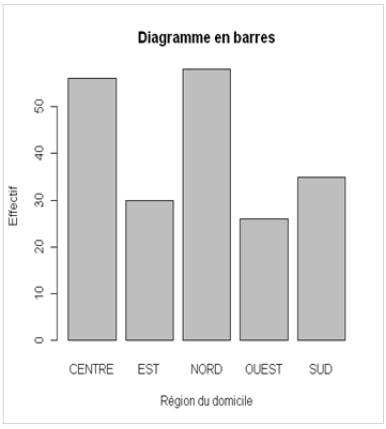
On peut obtenir le tableau des effectifs des catégories des variables nominales avec la fonction `table(NomDeVariable,...)`, qui est plus pratique que la mise en œuvre des tableaux croisés dynamiques de tableurs.

Les exemples des commandes figurent au-dessus des graphiques obtenus.

La commande `plot()` produit le graphique dans une nouvelle fenêtre graphique, avec des choix d'échelles par défaut, mais que l'on peut modifier en utilisant des paramètres additionnels.

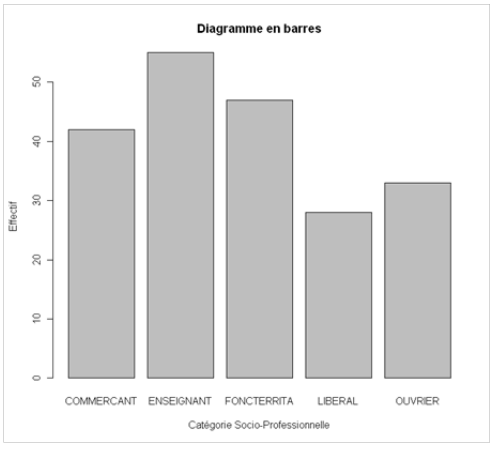
Ce choix par défaut est utile dans les séances d'initiation avec les élèves car cela permet de mieux se concentrer sur l'utilisation et l'interprétation des graphiques, en évitant dans un premier temps des difficultés dues à la construction fine du graphique.

```
plot(RegionDomicile, xlab = "Région
  du domicile", ylab = "Effectif",
  main = "Diagramme en barres")
```




```
# tri à plat de la variable
RegionDomicile
(table(RegionDomicile))
RegionDomicile
CENTRE    EST   NORD  OUEST    SUD
    56     30     58     26     35
```

```
plot(CSP, xlab = "Catégorie Socio-
  Professionnelle", ylab =
  "Effectif", main = "Diagramme en
  barres")
```



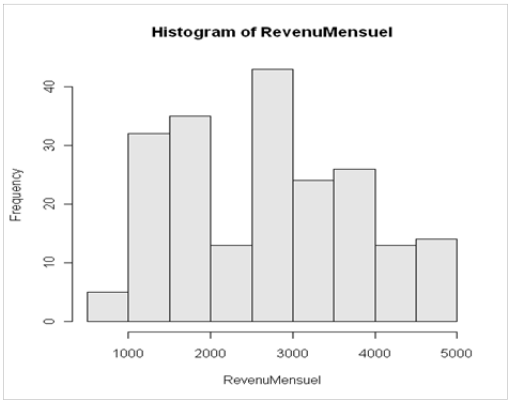

```
(table(CSP))
CSP
 COMMERCANT  ENSEIGNANT FONCTERRITA
         42          55          47
    LIBERAL     OUVRIER
         28          33
```

3.2.2 Description de distributions de variables quantitatives

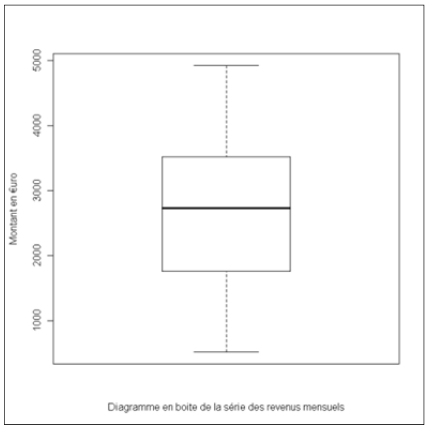
La fonction **`hist(NomDeVariable,...)`** produit un histogramme avec un découpage en classes et des étiquettes par défaut, mais que l'on peut paramétrer à volonté, comme le montre l'exemple ci-dessous.

La fonction **`boxplot(NomDeVariable,...)`** produit un diagramme en boîte (appelé aussi "boîte et moustaches" ou "boîte et pattes") de la totalité de la série des revenus mensuels.

```
# Histogramme de RevenuMensuel
hist(RevenuMensuel, breaks =
  seq(500, 5000, 500), right = F,
  col = "grey90")
```




```
boxplot(RevenuMensuel, ylab =
  "Montant en €uro", xlab =
  "Diagramme en boîte de la série
  des revenus mensuels")
```



Il est intéressant de remarquer que c'est l'histogramme qui permet de voir que la distribution observée n'est pas homogène et qu'il faut donc analyser l'effet des variables qualitatives, telles que *CSP* ou *RegionDomicile*, voire les combinaisons de modalités issues du croisement de ces deux variables.

3.3 Principaux outils de comparaisons

3.3.1 Comparer les distributions d'une variable quantitative observée dans différents "groupes"

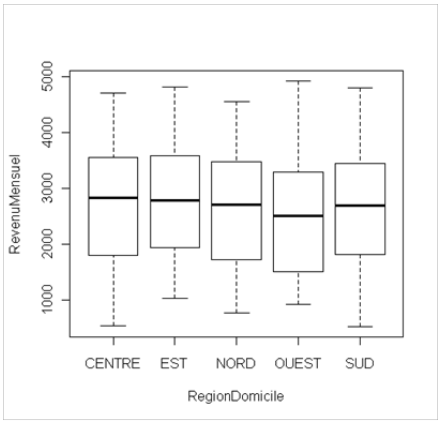
(i) Juxtaposer et comparer les diagrammes en boîtes des groupes ou les diagrammes-points

Croiser les variables *RevenuMensuel* et *RegionDomicile* consiste à représenter graphiquement (diagrammes en boîtes ou diagrammes-points) les séries observées (variable quantitative) en fonction des modalités de la variable qualitative.

C'est dans ce type de représentation que se révèlent toute la spécificité et l'efficacité de **R** pour l'analyse exploratoire, car ce type de juxtaposition des graphiques, en fonction d'une variable qualitative, n'est pas possible à réaliser avec un tableur.

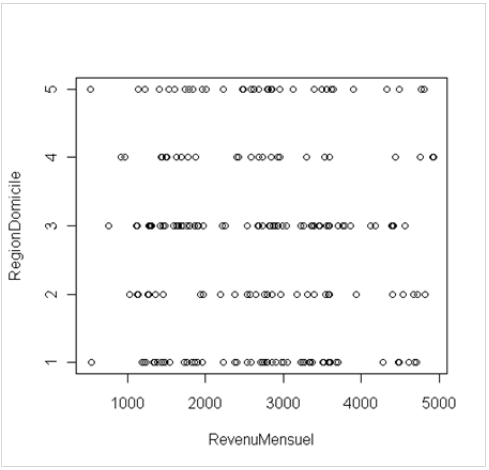
H. Raymondaud

```
# RevenuMensuel×RegionDomicile
boxplot(RegionDomicile ~
   RevenuMensuel, ylab =
   "RevenuMensuel", xlab =
   "RegionDomicile")
```




```
# par contre on obtient un autre
# graphique avec

plot(RevenuMensuel, RegionDomicile)
```

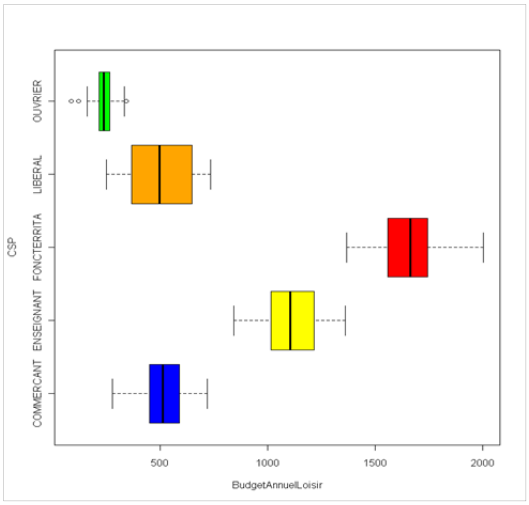


Les diagrammes-points sont une représentation non résumée (à la précision graphique près) de toutes les valeurs des séries, alors que les diagrammes en boîtes sont un résumé par des paramètres de rangs, les cinq quartiles (en admettant de nommer quartiles 0 et 4 le minimum et le maximum). Dans la version utilisée ici, le minimum est remplacé par la plus petite valeur de la série supérieure au premier quartile moins 1,5 fois l'intervalle interquartile et le maximum par la plus grande valeur de la série inférieure au troisième quartile plus 1,5 fois l'intervalle interquartile. Les comparaisons visuelles de groupes sont plus faciles avec les diagrammes en boîtes, alors que les diagrammes-points permettent de mieux visualiser les distributions observées de la variable sur les différents groupes.

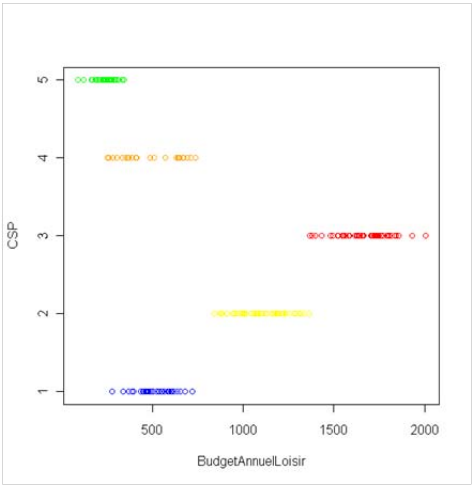
La simple utilisation de paramètres particuliers (**`horizontal, col`**) comme arguments de la fonction **`boxplot()`** permet de modifier la présentation des graphiques pour l'adapter à l'effet recherché, comme le montrent les deux graphiques ci-dessous, qui permettent de comparer le budget annuel loisir des différentes *CSP*.

La couleur rend l'identification des *CSP* et les comparaisons plus faciles. La juxtaposition des diagrammes en boîtes et des diagrammes-points permet de voir l'effet d'une distribution sur la forme de la boîte (position, dispersion, symétrie). Ce type de représentation n'est pas possible avec un tableur si ce n'est avec une programmation complexe (visual basic), hors de portée des classes de lycée.

```
# BudgetAnnuelLoisir×CSP
boxplot(BudgetAnnuelLoisir ~ CSP,
   horizontal = T, ylab = "CSP", xlab
   = "BudgetAnnuelLoisir", col =
   c("blue", "yellow", "red",
   "orange", "green"))
```




```
plot(BudgetAnnuelLoisir, CSP, col
   = c("blue", "yellow", "red",
   "orange", "green")[CSP])
```



Les commandes de **R** sont simples, la syntaxe est facile à comprendre, les possibilités d'amélioration des graphiques, à l'aide de paramètres, sont très nombreuses.

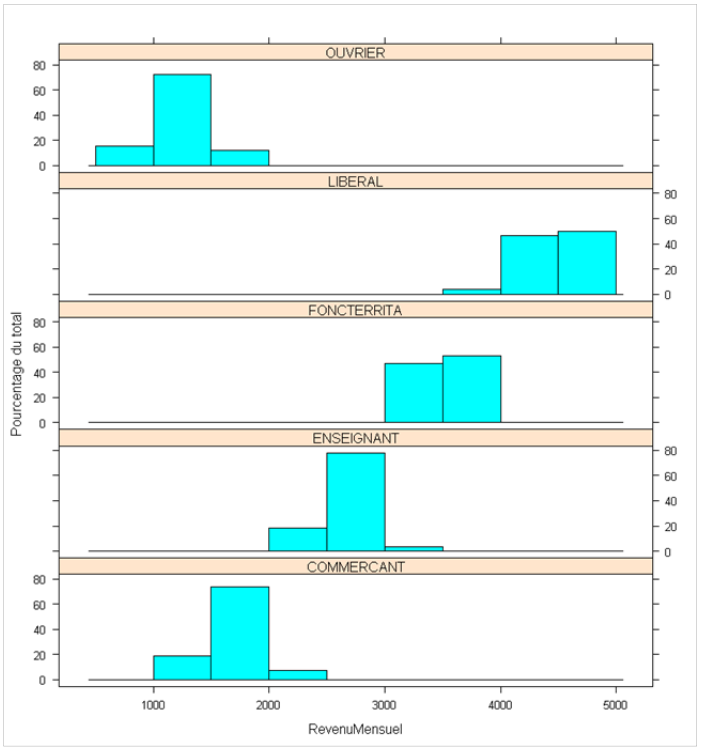
(ii) Juxtaposer et comparer les histogrammes

Nous avons vu un peu plus haut que l'histogramme de la série des revenus mensuels permettait de mettre en évidence une distribution non homogène. Il faudrait pouvoir réaliser les histogrammes par catégorie de *CSP*.

Cette juxtaposition d'histogrammes peut être réalisée manuellement mais c'est long et fastidieux car il faut paramétrer chacun des histogrammes de façon à ce que l'on puisse les comparer.

On peut réaliser automatiquement des histogrammes par catégorie d'une variable qualitative, en utilisant un "package" spécialement conçu pour faire des graphiques complexes, "lattice", qui doit être chargé dans **R** pour être utilisé. La commande est simple, comme le montre l'exemple de l'histogramme des séries de revenus mensuels par *CSP* :

```
require(lattice)
histogram(~ RevenuMensuel | CSP, layout = c(1, 5))
```



On met en évidence que c'est la variable CSP qui est responsable de la structure observée dans la distribution de la série. La position des distributions varie en fonction de la CSP.

3.3.2 Résumés numériques correspondant aux croisements variables quantitatives × variables qualitatives

Pour déterminer un ensemble de paramètres centraux et de tendance d'une série quantitative, pour chaque modalité d'une variable qualitative, on peut utiliser le package “Hmisc”.

Dans l'exemple ci-dessous, on crée une fonction **`moyetqu()`** qui calcule l'effectif, la moyenne, l'écart-type, les quantiles d'ordre 0%, 10%, 25%, 50%, 75%, 90%, 100% d'une série (les quantiles 0% et 100% désignant le min et le max). En utilisant cette fonction comme argument dans la fonction **`summary()`**, on obtient les valeurs calculées avec **`moyetqu()`**, pour la variable *RevenuMensuel*, par catégorie de CSP, présentées dans un tableau.

```
moyetyqu <- function (x, pquant = c(0, .10, .25, .50, .75, .90, 1)){
   moy <- mean(x)
   et <- sd(x)
   quant <- quantile(x, prob = pquant)
   c(Moyenne = moy, Ecartype = et, quant)
}
require(Hmisc)

summary(RevenuMensuel ~ CSP, fun = moyetyqu)
```

```
RevenuMensuel    N=205                  +          Quantiles                          +
+-------+-----------+---+--------+---------+----+------+-------+------+-------+------+----+
|       |           |N  |Moyenne |Ecartype |0%  |10%   |25%    |50%   |75%    |90%   |100%|
+-------+-----------+---+--------+---------+----+------+-------+------+-------+------+----+
|CSP    |COMMERCANT | 42|1747.881| 231.0901|1144|1436.2|1628.75|1780.0|1900.00|1969.9|2219|
|       |ENSEIGNANT | 55|2706.764| 214.0431|2227|2400.0|2576.00|2732.0|2852.50|2954.0|3170|
|       |FONCTERRITA| 47|3476.191| 205.4495|3004|3223.6|3338.00|3515.0|3592.50|3703.4|3940|
|       |LIBERAL    | 28|4528.679| 245.2382|3866|4244.3|4398.00|4513.5|4710.25|4807.3|4925|
|       |OUVRIER    | 33|1234.030| 262.3907| 520| 929.6|1123.00|1278.0|1444.00|1496.6|1587|
+-------+-----------+---+--------+---------+----+------+-------+------+-------+------+----+
|Overall|           |205|2698.488|1087.7611| 520|1319.8|1759.00|2732.0|3515.00|4398.0|4925|
+-------+-----------+---+--------+---------+----+------+-------+------+-------+------+----+
```

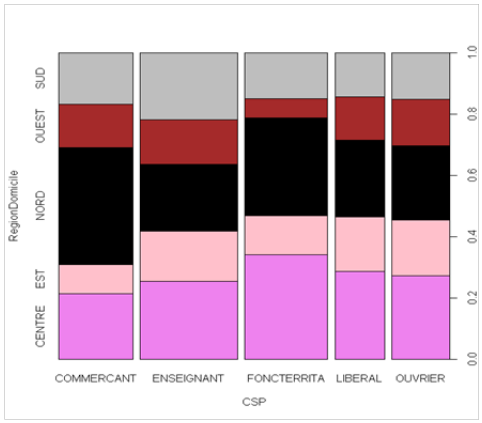
3.3.3 Étudier la relation entre deux variables qualitatives en établissant la distribution croisée de ces deux variables ou les distributions conditionnelles

Étudier l'association entre deux variables qualitatives par le biais d'une représentation graphique n'est pas toujours aisé. Il s'agit de faire comprendre aux étudiants qu'en réalité le statisticien a le choix entre trois représentations du tableau à double entrée, soit l'une ou l'autre des représentations des distributions conditionnelles ou la distribution des effectifs du tableau. L'exemple traité ci-dessous est celui de l'étude du tableau à deux entrées, représentant les effectifs des combinaisons des modalités de *RegionDomicile* et *CSP*.

La fonction **`table()`** permet d'obtenir le tableau des effectifs des combinaisons des modalités des deux variables.

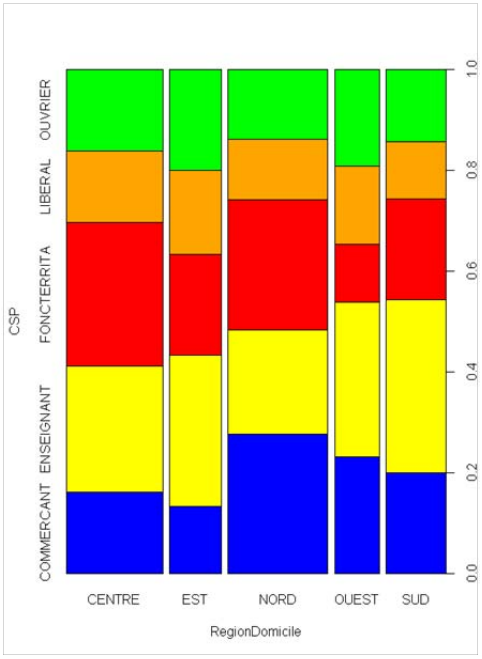
La fonction **`plot(variable1,variable2,…)`** permet d'obtenir la représentation graphique en barres superposées des fréquences relatives conditionnelles des deux variables. Il est intéressant de remarquer que la fonction **`plot()`** reconnaît la nature des variables et adapte automatiquement le type de graphique. On est très loin du fonctionnement d'un tableur...

```
# Tri croisé CSP×RegionDomicile
table(CSP, RegionDomicile)
                RegionDomicile
  CSP           CENTRE EST NORD OUEST SUD
    COMMERCANT       9   4   16     6   7
    ENSEIGNANT      14   9   12     8  12
    FONCTERRITA     16   6   15     3   7
    LIBERAL          8   5    7     4   4
    OUVRIER          9   6    8     5   5
plot(CSP, RegionDomicile, col =
  c("violet", "pink", "black",
  "brown", "grey"), xlab = "CSP",
  ylab = "RegionDomicile")
```



Attention l'impression en noir et blanc du graphique ci-dessus sera peu lisible

```
plot(RegionDomicile, CSP, col =
  c("blue", "yellow", "red",
  "orange", "green"), ylab =
  "CSP", xlab =
  "RegionDomicile")
```



Attention l'impression en noir et blanc du graphique ci-dessus sera illisible

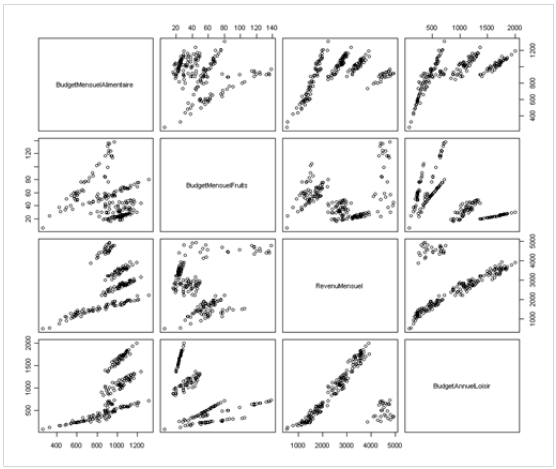
3.3.4 Étudier la relation entre deux variables quantitatives en établissant des nuages de points

Dans un premier temps il s'agit de choisir les couples de variables quantitatives que l'on veut croiser. Si l'on n'a pas de raison de privilégier certains couples, **R**, simplement avec la commande **`plot(variable1,variable2,…)`**, nous offre la possibilité de présenter, en une seule fois, tous les croisements 2 à 2, dans une "matrice" de nuages de points.

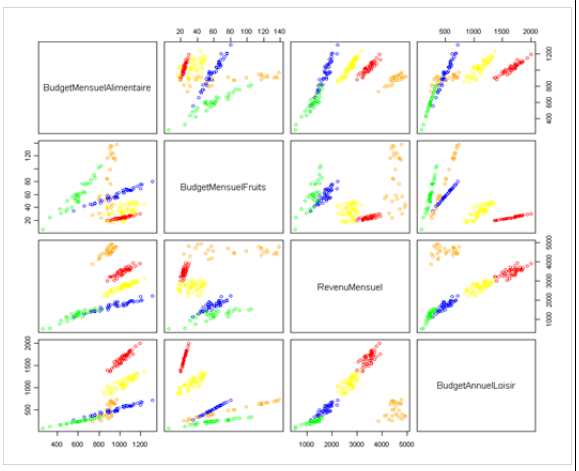
L'utilisation des couleurs pour marquer les *CSP* permet de repérer les facteurs des structures dans les nuages de points. Marquer les *CSP* par la couleur revient à faire un croisement par une troisième variable qualitative.

On peut ainsi rechercher quel facteur est à l'origine des structures observées et cela de manière rapide est efficace.

```
plot(habit[,3:6])
# porte sur les colonnes 3 à 6 de
   la dataframe habit
```

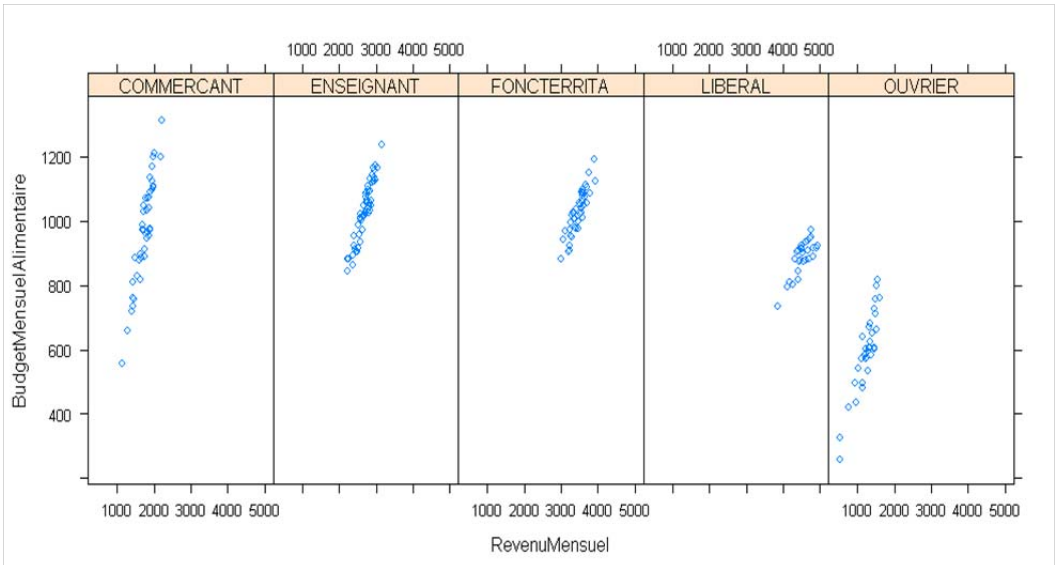



```
Ou encore mieux, en marquant les
CSP par des couleurs :
plot(habit[,3:6], col = c("blue",
   "yellow", "red", "orange",
   "green")[CSP])
```



Le package "lattice" propose des fonctions permettant de répartir automatiquement les nuages en fonction de la *CSP*, ce qui offre une lecture encore facilitée, comme cela est visible dans la représentation ci-dessous, avec la fonction **`xyplot()`**.

```
require(lattice)
xyplot(BudgetMensuelAlimentaire ~ RevenuMensuel | CSP)
```



Là encore on est très loin des possibilités d'un tableur...

Avec des étudiants de BTSA, la séquence a été réalisée en deux séances de deux heures, comprenant quelques rappels sur les méthodes et outils de la statistique descriptive.

4 Conclusion

J'utilise **R** pour trois activités différentes : l'illustration de mes cours, le traitement de données de l'expérimentation agricole et la formation des élèves de BTS au traitement des données. Dans ces trois activités **R** a progressivement remplacé l'utilisation de logiciels spécifiques et les tableurs.

Les exemples de cet article ont montré quelques différences d'importance entre un tableur et un véritable logiciel de traitement statistique. Dans la pratique du traitement des données l'efficacité de **R** facilite les analyses et offre un panel étendu de méthodes.

Les fonctions graphiques qui reconnaissent les types de variables et produisent les graphiques adaptés, la possibilité d'utiliser des variables qualitatives pour analyser et comparer des groupes, des possibilités de requêtes permettant de travailler sur une partie des données sans être obligé de manipuler des tableaux de données, la possibilité de construire ses propres fonctions, sont quelques-unes des spécificités d'un véritable logiciel statistique.

Un atout d'importance de **R** est la grande variété de méthodes offertes dans les packages développés par les spécialistes des méthodes proposées.

Après avoir longtemps utilisé les tableurs, j'ai choisi **R** comme outil privilégié pour l'apprentissage des traitements statistiques avec les élèves. En effet, les difficultés rencontrées lors des nombreuses manipulations des données, le temps important nécessaire lorsqu'il s'agit de faire des traitements par groupe, la difficulté de gérer de grands lots de données, rendent les traitements statistiques avec un tableur longs et laborieux avec des élèves parfois peu familiarisés avec l'outil informatique.

R, avec son langage facile à comprendre et à lire, permet de construire des progressions que les élèves peuvent suivre pour réaliser rapidement et facilement les traitements demandés et ainsi mieux se concentrer sur l'analyse et l'interprétation des résultats et la recherche de traitements complémentaires.

Il me semble donc que **R** peut avoir une place privilégiée au lycée pour l'apprentissage et la mise en œuvre des méthodes statistiques.

En prolongement aux méthodes de l'analyse exploratoire des données, **R** offre des possibilités étendues dans le domaine des probabilités et de l'inférence statistique.

Enfin, son langage permet de concevoir des simulations simples ou complexes, en probabilité ou en inférence, ce qui en fait un outil de choix pour mettre en œuvre les simulations et l'algorithmique présents dans les programmes de la seconde à la terminale. C'est à ce titre que j'ai proposé d'introduire **R** dans le document ressource des nouveaux programmes de terminales S et ES.

Références

[1] Lafaye De Micheaux, P., R. Drouilhet et B. Liquet (2010), *Le logiciel **R** – Maîtriser le langage, effectuer des analyses statistiques*, Springer.

[2] Millot, G. (2009), *Comprendre et réaliser des tests statistiques à l'aide de **R** – Manuel pour les débutants*, De Boeck.

[3] Husson, Fr., S. Lê et J. Pagès (2009), *Analyse de données avec **R,*** Presses Universitaires de Rennes.

[4] Cornillon, P.-A. (2010), *Régression avec **R***, Springer.

[5] Cornillon, P.-A., A. Guyader, Fr. Husson *et al.* (2010), *Statistique avec **R*** (2e édition), Presses Universitaires de Rennes.

[6] Sarkar, D. (2008), *Lattice, Multivariate Data Visualization with **R***, Springer.

[7] Bertrand, Fr. (2010), *Initiation aux statistiques avec **R** ; cours, exemples, exercices et problèmes corrigés*, Licence 3, Master 1, Ecoles d'ingénieur, Dunod.

[8] Robert, Chr. P. et G. Casella (2011), *Méthodes de Monte-Carlo avec **R***, Springer.

[9] Enseignement de statistique en biologie, Université de Lyon 1 : http://pbil.univ-lyon1.fr/R/enseignement.html

[10] Semin-R, groupe d'utilisateurs de R : http://rug.mnhn.fr/semin-r/

[11] Hoaglin, D. C., Fr. Mosteller, and J. W. Tukey (1982), *Understanding Robust and Exploratory Data Analysis*, Wiley, Series in Probability and Statistics.

[12] Document ressource des nouveaux programmes de terminales S et ES : téléchargeable sur le site Eduscol, http://eduscol.education.fr/cid45766/ressources-pour-faire-la-classe-au-college-et-au-lycee.html